

NIST
PUBLICATIONS

REFERENCE

NISTIR 7020

Studies of Fingerprint Matching Using the NIST Verification Test Bed (VTB)

***Charles L. Wilson^a
Craig I. Watson^a
Michael D. Garriss^a
Austin Hicklin^b***

**^aU. S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Gaithersburg, MD 20899**

^bMitretek Systems

QC
100
.U56
NO.7020
2003

NIST

**National Institute of Standards
and Technology
Technology Administration
U.S. Department of Commerce**

***Studies of Fingerprint Matching Using the
NIST Verification Test Bed (VTB)***

***Charles L. Wilson^a
Craig I. Watson^a
Michael D. Garris^a
Austin Hicklin^b***

**^aU. S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Gaithersburg, MD 20899**

^bMitretek Systems

July 7, 2003



**U.S. DEPARTMENT OF COMMERCE
Donald L. Evans, Secretary**

**TECHNOLOGY ADMINISTRATION
Phillip J. Bond, Under Secretary for Technology**

**NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
Arden L. Bement, Jr., Director**

TABLE OF CONTENTS

Abstract	1
1. INTRODUCTION.....	1
1.1 Brief History of Biometrics at NIST	2
1.2 Change in Focus as of 9-11	2
1.2.1 USA PATRIOT Act Requirements	3
1.2.2 Border Security Act Requirements	3
1.2.3 303A Report	3
1.3 Need for the VTB	4
1.4 Report Organization	5
2. VTB DESCRIPTION.....	5
2.1 Hardware Description	5
2.2 Software Description.....	6
2.2.1 NIST Fingerprint Image Software.....	6
2.2.2 Four-Finger Plain Segmenter	6
2.2.3 Bozorth98 Fingerprint Minutiae Matcher	9
2.2.4 Scoring Software	10
3. VTB DATA REPOSITORIES	10
3.1 NIST Special Database 14 (SD14).....	11
3.2 NIST Special Database 24 (SD24).....	12
3.3 NIST Special Database 29 (SD29).....	13
3.4 Immigration and Naturalization Service Recidivist Database	14
3.4.1 DHS 2-Finger Images (DHS2).....	14
3.4.1.1 Matcher-Based Quality Control	15
3.4.2 DHS 10-Finger Images (DHS10).....	16
3.4.3 DHS Consolidation Set of 10-Finger Images (DHS10-C).....	17
3.5 Department of State Mexican Visa Database (DOS)	18
3.6 Texas Department of Public Safety Database (TXDPS).....	19
4. EVALUATION FRAMEWORK.....	20
4.1 Terminology and Definitions	20
4.2 Verification vs. Identification	20
5. STUDIES AND RESULTS	21
5.1 Overview of Studies	21
5.1.1 Small-Scale Studies.....	21
5.1.2 Large Scale Studies	22
5.1.3 Other Studies	22
5.2 Inked, Rolled Impression Verification Study with SD14	23
5.3 Live-Scan, Plain Impression Verification Study with SD24.....	24
5.4 Inked, Rolled vs. Plain Impression Verification Study with SD29.....	26
5.5 Live-Scan, Rolled vs. Plain Impression Verification Study with DHS10-C	32
5.5.1 DHS10 Consolidation	32
5.5.2 DHS10-C Results	32
5.6 Large Scale Live-Scan Verification Study with DHS2.....	35
5.7 Large Scale Live-Scan Verification Study with DOS.....	38
5.8 Large Scale Inked Verification Study with DHS10	40

5.9	Large Scale Inked Verification Study with TXDPS	42
5.10	Large Scale Identification Study with DHS2	44
5.11	Large Scale Identification Study with DOS	46
5.12	Fusion of Results from Multiple Fingerprints.....	48
5.12.1	Score-Based Fusion Using SD29	48
5.12.2	Score-Based Fusion Using DHS10-C	55
5.12.3	Rank and Score-Based Fusion Using DHS2.....	58
5.13	Person Variation Study with DHS2	61
6.	Implications of Metadata.....	73
6.1	DHS2 Metadata Study.....	74
6.1.1	Nonstationary Results Observed	74
6.1.2	Metadata Analyzed.....	75
6.1.3	DHS2 Metadata Study Summary	83
6.2	Concept for a Fingerprint Experiment Manager	83
7.	Conclusions	83
7.1	Critical Test Parameters	84
7.2	Small Sample Test Conclusions.....	84
7.3	Large Sample Test Conclusions.....	85
	REFERENCES.....	89
	APPENDIX A. Matcher Comparisons to the VTB	92

LIST OF FIGURES

Figure 1.	FBI's FD-249 tenprint card for criminal cases	8
Figure 2.	Cropped tenprint card filled with fingerprints	8
Figure 3.	SD14 Verification Study – Comparison of right and left thumbs and index fingers....	24
Figure 4.	SD24 Verification Study – Comparison of thumb, index, middle, ring, and little fingers	25
Figure 5.	SD29 Thumb Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled.....	27
Figure 6.	SD29 Index Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled.....	28
Figure 7.	SD29 Middle Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled.....	29
Figure 8.	SD29 Ring Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled.....	30
Figure 9.	SD29 Little Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled.....	31
Figure 10.	DHS10-C Right Thumb Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled.....	33
Figure 11.	DHS10-C Right Index Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled	34
Figure 12.	DHS2 Plain-to-Plain Right Index Finger Large Scale Verification –Multi-Trial ROC	36
Figure 13.	DHS2 Right Index Finger Large Scale Verification –Submatrix variation of Multi-Trial ROC.....	37

Figure 14. DOS Index Finger Large Scale Verification –Multi-Trial ROC	38
Figure 15. DOS Index Finger Large Scale Verification – Submatrix variation of Multi-Trial ROC.....	40
Figure 16. DHS10 Right Thumb vs. Right Index Finger Large Scale Verification –Multi-Trial ROC (Plain-to-Rolled)	41
Figure 17. TXDPS Thumb vs. Index Finger Large Scale Verification –Multi-Trial ROC.....	43
Figure 18. DHS2 Right & Left Index Finger Identification - % Correct by blocks of 100 people	45
Figure 19. DHS2 Large Scale Identification – Comparison of right and left index fingers	46
Figure 20. DOS Right & Left Index Finger Identification - % Correct by blocks of 100 people.....	47
Figure 21. DOS Large Scale Identification – Comparison of right and left index fingers	48
Figure 22. SD29 Combined Thumb & Index Finger Scores.....	50
Figure 23. SD29 Combined Thumb & Index Finger Verification	51
Figure 24. SD29 Combined Right & Left Index Finger Scores.....	53
Figure 25. SD29 Combined Right & Left Index Finger Verification	54
Figure 26. DHS10-C Combined Right Thumb & Right Index Finger Verification	56
Figure 27. DHS10-C Combined Right & Left Index Finger Verification	57
Figure 28. DHS2 Combining Right and Left Index Finger Identification – Comparing combined results with right index finger alone.....	59
Figure 29. Identification Using Right Index Finger Alone	60
Figure 30. Right Index Fingers Scoring in Top-100.....	60
Figure 31. Identification Combining Left with Right Index Finger Scores	61
Figure 32. Composition of Similarity Matrix for Person Variation Study.....	62
Figure 33. Good Quality Fingerprints from DHS2 Person 9	64
Figure 34. Intra-Person Similarity Submatrix for Good Quality Case 1.....	65
Figure 35. Poor Quality Fingerprints from DHS2 Person 7.....	66
Figure 36. Comparison of Intra-Person Matcher Scores Between a Good and Poor Quality Case	67
Figure 37. Good Quality Fingerprints from DHS2 Person 24	68
Figure 38. Comparison of Intra-Person Matcher Scores Between Two Good Quality Cases	69
Figure 39. Distribution of Inter-Person Matcher Scores Between Two Good Quality Cases	70
Figure 40. Mean Submatrix Matcher Scores – Comparing intra-person to inter-person submatrices.....	71
Figure 41. Standard Deviation of Submatrix Matcher Scores – Comparing intra-person to inter-person submatrices	72
Figure 42. Correlation Plot of Mean vs. Standard Deviation from Submatrix Matcher Scores - Comparing intra-person to inter-person submatrices.....	73
Figure 43. DHS2 Right Index Finger Large Scale Verification – <i>Procedurally-Selected</i> Submatrix variation of Multi-Trial ROC	75
Figure 44. Frequency of Encounters across Time.....	77
Figure 45. Quality of Images Captured across Time	78
Figure 46. Gender across Time	79
Figure 47. Frequency of Encounters across Capture Location	80
Figure 48. Quality of Images across Capture Location.....	81
Figure 49. Gender across Capture Location.....	82
Figure 50. Comparison of Plain-to-Plain Results from SD29, DHS2, & DOS.....	86

Figure 51. Comparison of Results from SD24, DHS10, & TXDPS	87
Figure 52. ROC curves for three algorithms (VTB and two commercial fingerprint matchers) and four data sets.....	93

LIST OF TABLES

Table 1. SD14 Results.....	24
Table 2. SD24 Results	25
Table 3. SD29 Results.....	31
Table 4. DHS10-C Results.....	34
Table 5. DHS2 Results.....	36
Table 6. DOS Results	39
Table 7. DHS10 Right Thumb vs. Right Index Results	41
Table 8. TXDPS Thumb vs. Index Results	43
Table 9. SD29 Score-Based Thumb and Index Finger Fusion Results.....	51
Table 10. Actual and Ideal Score-Based Fusion for Index Fingers and Thumbs.....	52
Table 11. SD29 Score-Based Right and Left Index Finger Fusion Results	54
Table 12. Actual and Ideal Score-Based Fusion for Index Fingers	55
Table 13. DHS10-C Score-Based Right Thumb and Right Index Finger Fusion Results	56
Table 14. DHS10-C Score-Based Right and Left Index Finger Fusion Results.....	57
Table 15. DHS2 Metadata.....	76
Table 16. Correlation Table of Metadata Factors	82
Table 17. TAR for three algorithms and four data sets at 1% FAR.....	93

Studies of Fingerprint Matching Using the NIST Verification Test Bed (VTB)

Charles L. Wilson^{*}, Craig I. Watson^{*}, Michael D. Garriss^{*}, & Austin Hicklin[†]

Abstract

A series of fingerprint matching studies have been conducted on an experimental laboratory system called the Verification Test Bed (VTB). The VTB is a collection of commercial off the shelf (COTS) computer hardware, an open-source operating system, and a suite of public domain application software. Results are presented that compare various sources of fingerprints, assess the image quality of fingerprints by analyzing the matcher scores for inked and live-scan impressions, and study the trade-offs of matching rolled fingerprints with plain impressions. Database size in these studies range from 216 to ~600,000 people. Performance statistics are primarily reported for single-finger matching; however, results from two different approaches to two-finger fusion matching are also presented. At a false accept rate (FAR) of 1%, the best two-finger true accept rate (TAR) is 99% while the worst single-finger TAR is 71%. This report illustrates the wide range of image types and quality that exist in government fingerprint databases and the effect this variability has on the accuracy of matching using a single algorithm.

An appendix compares the VTB matcher to two commercial fingerprint systems and concludes that the performance of the VTB is very similar to commercial verification systems currently on the market. This further confirms that data quality uniformity is of paramount importance in the evaluation of fingerprint biometrics.

Keywords: fingerprint, identification, matching, system evaluation, verification

1. INTRODUCTION

This report documents a series of fingerprint matching studies conducted on an experimental laboratory system called the Verification Test Bed or VTB. These studies span approximately eight months of research (from September 2002 to April 2003) in the Information Access Division's Image Group at the National Institute of Standards and Technology (NIST). The VTB is a collection of commercial off the shelf (COTS) computer hardware, an open-source operating system, and a suite of public domain application software, unlike most fingerprint

^{*} Mr. Wilson, Mr. Watson, and Mr. Garriss are employees with the National Institute of Standards and Technology in Gaithersburg, MD.

[†] Mr. Hicklin is an employee of the Mitretek Systems in Falls Church, VA.

matchers, which are expensive to obtain, and require specialized hardware. The VTB was developed to be a reference matcher that can provide a performance baseline for future analyses of fingerprint matchers, as well as comparative analysis of different sets of fingerprint data. As will be discussed in this report, this “open” system has proven critical to our mission.

1.1 Brief History of Biometrics at NIST

NIST has a long history of involvement in biometric research and biometric standards development. For over 30 years, NIST has collaborated with the Federal Bureau of Investigation (FBI) in the area of automated fingerprint recognition. Researchers at NIST (then the National Bureau of Standards (NBS)) began work on the first version of the FBI's Automated Fingerprint Identification System (AFIS) system back in the late 1960's. Over the years, NIST has conducted fingerprint research, developed fingerprint identification technology and data exchange standards, developed methods for measuring the quality and performance of fingerprint scanners and imaging systems, and produced databases containing fingerprint images for public distribution [1]-[30].

The Image Group sponsored one of the most influential biometric standards in the law enforcement community. This is the ANSI/NIST-ITL 1-2000 "Data Format for the Interchange of Fingerprint, Facial, Scar Mark & Tattoo (SMT) Information" standard [30]. This standard (referred to as ANSI/NIST 2000) defines a common file format, available to law enforcement agencies in the U.S. since 1986 [9], for the electronic exchange of fingerprint images and related data. Today, it supports other types of images as well, including palmprints, mugshots, scars, and tattoos. This standard has been adopted by all major law enforcement agencies in the U.S., including the FBI, and has strong support and use internationally.

More recently, the Image Group has run a series of large scale face recognition system tests called FRVT2000 [31] and FRVT2002 [32]. Conducting these technology evaluations requires collection and publication of large volumes of data as well as development of scoring technology for the computation of performance statistics. As a result, NIST has significant experience and expertise in managing and analyzing large repositories of biometric data, and it has developed a testing framework and protocol called the HumanID Evaluation Framework (HEF) [33] for evaluating the performance of biometric systems.

It is accurate to say that NIST has a long history in biometrics with emphasis on law enforcement fingerprint applications and standards. Based on this experience, it was not too surprising that Congress included NIST in its legislative response to the terrorist attacks on September 11, 2001 (9-11).

1.2 Change in Focus as of 9-11

Since 9-11, the phrase, “Everything has changed,” has been frequently stated. This is no less true for the Image Group at NIST. Within a couple of months, new initiatives were started that redirected work focused on law enforcement to new work focused on border control.

On the heels of 9-11 came several pieces of congressional legislation which directly cited participation and contribution from NIST in the area of biometric standards development. These include the USA PATRIOT Act and the Enhanced Border Security and Visa Entry Reform Act.

Both of these acts specify requirements for interoperable biometric systems that are being developed by the Department of Homeland Security (DHS) and the Department of State (DOS). Specifically the requirements are:

1.2.1 USA PATRIOT Act Requirements

The USA PATRIOT Act, in section 403(c)(1), as amended by the Enhanced Border Security and Visa Entry Reform Act, directs that the Attorney General and the Secretary of State jointly, through NIST “shall [] develop and certify a technology standard, including appropriate biometric identifier standards, that can be used to verify the identity of persons applying for a United States visa or such persons seeking to enter the United States pursuant to a visa for the purposes of conducting background checks, confirming identity, and ensuring that a person has not received a visa under a different name.”

1.2.2 Border Security Act Requirements

The Enhanced Border Security and Visa Entry Reform Act states in section 202(a)(3) that “In the development and implementation of the data system under this subsection, the President shall consult with the Director of the National Institute of Standards and Technology (NIST) and any such other agency as may be deemed appropriate.”

In addition section 202(a)(4)(A) states that “The data system developed and implemented under this subsection, and the databases referred to in paragraph (2), shall utilize the technology standard established pursuant to section 403(c) of the USA PATRIOT Act ...”

These standards apply to visas documents issued by the US government. A visa waiver country is required by section 303(c)(1) “to issue to its nationals machine-readable passports that are tamper-resistant and incorporate biometric and document authentication identifiers that comply with applicable biometric and document identifying standards established by the International Civil Aviation Organization.”

1.2.3 303A Report

Previously, the report titled “Use of Technology Standards and Interoperable Databases With Machine-Readable, Tamper-Resistant Travel Documents” was submitted to the Congress jointly by the Attorney General, Secretary of State, and NIST [34]. (This report is informally referred to as the 303A Report.) It discusses measurements of the accuracy of both face and fingerprints as they relate to U.S. border entry and exit. The detailed face recognition results are documented in the FRVT 2002 report [32]. The fingerprint results in the report were calculated using the data and evaluation methods discussed in this report.

The report submitted to Congress concluded that:

NIST has determined that face and fingerprints are the only biometrics available with large enough operational databases for testing at this time. Both technologies are mature. To properly certify any biometric, extensive tests must be performed using databases containing at least 100,000 subjects. Such databases have been acquired from NIST, FBI, DHS, DOS, and the Texas Department of Public Safety (DPS) to perform the required testing.

Results from fingerprint testing based on a Mitretek study, and NIST testing using SD24, and a sampling of DHS data have been analyzed. To perform background identifications, ten plain image impressions should be used for enrollment and retention. As described in the "FBI IAFIS Accuracy" section of the 303A Report, Mitretek recommends a minimum of four plain finger impressions for background searches. With the live-scan fingerprint scanners currently available, the additional time required to capture the additional six fingers will be insignificant.

Results show fingerprint matching to be accurate. Verification can be performed on single fingers with 90% accuracy at a false accept rate of 1%.[‡] Single finger identification can provide 95% accuracy for a gallery size of 500. The identification rate drops to 90% for a gallery size of 10,000 and to 86% for a gallery size of 100,000[§]. This test illustrates the difficult nature of accurate database searches using a single fingerprint. High accuracy searching of a database of 1 million subjects or greater will require more than one finger whether the FBI's IAFIS is used or not.

Results indicate that single fingerprints provide approximately the same verification accuracy as face. For facial recognition, the best packages available (based on FRVT 2002) provide a 90% probability of true verification with a 1% probability of false verification. This makes face recognition an excellent choice as an alternative to fingerprints for verification and for situations where fingerprints are not available and where high quality face images with good illumination control similar to those taken using the DOS visa protocol are available.

Under less constrained outdoor conditions face recognition accuracy for the best system falls to 47%. For identification the best available face recognition technology identification can provide 90% accuracy for a gallery size of 100. The identification rate drops to 83% for a gallery size of 1,000 and to 77% for a gallery size of 10,000. These numbers demonstrate that for identification, fingerprints are the preferred technology. However, not all subjects can be easily fingerprinted with existing technology resulting in a 2% failure to acquire rate.

Furthermore, within the intelligence community, facial data is often the only biometric data that has been and is currently being captured. Based on these considerations, our measurements indicate that a dual biometric system including two fingerprint images and a face image may be needed to meet projected system requirements for verification. Each fingerprint and the facial image should require 10 kilobytes or less of storage apiece. Therefore, a card capable of storing two fingerprints and one face image will require a 32K-byte chip to fulfill these requirements.

The body of this VTB report explains in more detail the fingerprint accuracy measurements used to draw these conclusions.

1.3 Need for the VTB

Given the severity and significance of the events of 9-11, agencies responsible for securing our country's borders are more intent than ever before to improve and integrate their biometric systems. As a result, there has been an unprecedented level of cooperation between these

[‡] These results are shown in this report using the DHS2 results in Figure 12.

[§] Similar results are shown in this report using the DHS2 results in Figure 19.

agencies and NIST. This has resulted in the delivery of prototype production systems and, even more significantly, the exchange of large working repositories of biometric data.

To carry out the legislated requirements of Congress, and to begin processing and analyzing these large repositories of data, NIST required a versatile *open* system for conducting applied research. The VTB was designed for the following purposes:

- To develop fingerprint evaluation methods and protocols and for evaluating baseline technology.
- To provide a large computation capacity for conducting fingerprint matches.
- To segment four-finger plain impressions so that rolled vs. plain studies can be conducted.
- To build large databases and conduct automated data quality checks to create repositories for use in future evaluations and on prototype production systems.

The VTB is being compared against other matchers in ongoing studies: its role here is to serve a minimum standard baseline for fingerprint matcher performance, and to allow comparative analysis of different types of fingerprint data.

1.4 Report Organization

The remainder of this report is devoted to documenting the VTB and a series of experiments conducted on it. Section 2 provides an overview of the VTB including hardware and software descriptions. Section 3 documents the various repositories of fingerprints that have been analyzed on the VTB. Section 4 presents an evaluation framework and defines key terminology used by NIST in its performance evaluations. Section 5 presents a lengthy series of studies and results. Section 6 examines the use of metadata. Finally, Section 7 draws conclusions.

2. VTB DESCRIPTION

The VTB is a system comprised of a collection of COTS hardware, an open-source operating system, and public domain software. A general description of what constitutes the VTB is presented in this section.

2.1 Hardware Description

The VTB is currently comprised of 16 dual-processor personal computers. All nodes are equally equipped with the following hardware:

- Dual 1.8Mhz Intel Xeon Processors with 512K Cache
- 400 MHz system bus
- 1 GB PC800 memory 400MHz ECC

- 64bit Gigabit Network card
- 64bit SCSI adapter card
- External IDE RAID with SCSI interface
 - 700GB capacity
 - 8-120GB ATA100, 7200RPM drives
 - Raid level 5 with 1 hot spare

2.2 Software Description

In addition to the Linux operating system (Red Hat Linux 7.2**), a suite of NIST application software was installed on each VTB node.

2.2.1 NIST Fingerprint Image Software

The NIST Fingerprint Image Software (NFIS) [35] provides many of the fingerprint capabilities required by the VTB. NFIS is a public domain source code distribution organized into four major packages:

1. PCASYS (Pattern Classification Automation SYstem) is a neural network based fingerprint pattern classification system;
2. MINDTCT (MINutiae DeTeCTOR) is a fingerprint minutiae detector;
3. AN2K (ANSI/NIST 2000) is a reference implementation of the ANSI/NIST 2000 standard; and
4. IMGTOOLS (IMaGe TOOLS) is a collection of image utilities, including encoders and decoders for Baseline and Lossless JPEG and the FBI's Wavelet Scalar Quantization (WSQ) specification.

NFIS is essential to the VTB. Fingerprint image files on the VTB are formatted according to ANSI/NIST 2000 and are compressed using WSQ. Minutiae are extracted from fingerprint images using MINDTCT. Note that PCASYS is not currently used in the VTB.

2.2.2 Four-Finger Plain Segmenter

A key issue to be addressed when considering next generation border control systems is the effect (if any) on searching legacy repositories of rolled impressions with plain impressions

** Specific hardware and software products identified in this report were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

A strategy was developed based on the fact that a standard tenprint card contains both a complete set of ten rolled finger impressions and also a corresponding set of plain impressions. Figure 1 shows a blank tenprint card (the FBI's FD-249). Boxes numbered 1–10 are designed to hold rolled impressions, while the bottom row of four boxes is designed to hold plain impressions. (Note there are many forms of fingerprint cards, but all contain fingerprint impressions in generally the same locations on the form.)

7

Figure 1. FBI's FD-249 tenprint card for criminal cases

Figure 2 shows a slightly different FBI FD-249 filled in with fingerprints. There are several interesting and important things to notice in this figure. First, the top portion of the card has been cropped, making the remaining fingerprint images anonymous. Now compare the rolled impressions in boxes 1–10 with the plain impressions in the bottom row of four boxes. The left-most box and right-most box at the bottom of the form each contain an impression of 4 fingers (index, middle, ring, and little) captured simultaneously, while the two middle boxes contain single plain impressions of the person's thumbs. From this example, one can see that a tenprint card contains two full sets of fingers, one rolled and one plain.



Figure 2. Cropped tenprint card filled with fingerprints

If one were to extract both sets of fingerprints from a tenprint card, then plain versus rolled studies could be conducted. The greatest challenge with this is that two groups (left and right hand) of four fingers (index, middle, ring, and little) are imprinted in a single box on the card. As seen in Figure 2, there is one box for the four fingers from the right hand (bottom right), and a second box for the four fingers from the left hand (bottom left).

A segmenter was designed to automatically extract plain impressions from an image of a four-finger plain impression box. Commercial segmenters are available, but there were unique requirements that made it necessary for NIST to develop its own technology. Commercial products have been designed specifically to process live-scan images, and they have difficulty

handling artifacts in the image such as handwriting, which is common on scanned images of tenprint cards. Commercial products are designed primarily to maximize yield with little feedback to automatically reject questionable segmentations. The NIST segmenter is carefully designed to find a compromise between maximum yield and accurate automatic results. It was anticipated that as many as one million card images would be processed, so complete automation with no manual interaction or verification was critical.

The NIST segmenter uses a down-sampled binary version of the four-finger plain image. A search is made, which includes rotation, for the best fit of four black ridges (fingers) and three white valleys (space between fingers) and if a sufficient fit cannot be found the plains are rejected and not used. After finding all four fingers, the fingertips are isolated by a window, sized just large enough to enclose the fingertip. If any errors occur while trying to isolate the fingertips, or if the final windows do not meet minimum size requirements, the plains are rejected. Finally, each fingerprint is copied from the original four-finger image, without removing rotation, into a new image with white background. Therefore, any pixels not filled by the copy are set to white.

If plain impressions are rejected for any reason, then the entire card is removed from the resulting repository. Using this approach with the NIST segmenter, about 50% of all cards processed are rejected. The high rejection rate is offset by the fact that the remaining results are highly accurate, and no human interaction is required to build the repository.

2.2.3 Bozorth98 Fingerprint Minutiae Matcher

The VTB detects and reports minutiae in a fingerprint image using MINDTCT distributed in NFIS [35]. Minutiae are points in a finger's friction skin where ridges end (called a *ridge ending*) or split (called a *ridge bifurcation*). These features are represented in their most fundamental form as an ordered triple (x, y, θ) ; where (x, y) is a minutia's Cartesian coordinate location, and θ designates the orientation of local ridge flow. Once minutiae are extracted, two different finger image impressions can be compared to each other by matching their corresponding sets of (x, y, θ) values.

The VTB uses a matcher algorithm referred to as the Bozorth98 matcher, which was chosen as the best available fingerprint matcher for which the algorithm and source code were readily available. The Bozorth98 matcher was developed and implemented by Alan Bozorth of the FBI. The algorithm, developed in 1993-95, was designed to match two sets of (x, y, θ) values in such a way as to be rotationally invariant. This capability enables the algorithm to match two fingerprints without the need to first compensate for the fact that the fingerprints may have been captured at different orientations.

To accomplish this, the algorithm transforms each fingerprint's set of (x, y, θ) values into a specialized rotationally invariant graph. To compute a match score between two fingerprints, the algorithm iteratively searches between both fingers' graphs for subsets (or subgraphs) that are *compatible*, i.e. coordinate locations and orientations of the minutiae represented within the subgraphs are similar enough to each other based on heuristically defined tolerances. The more nodes contained within a compatible subgraph, the higher the accumulated match score. The

more subgraphs that are compatible between the two fingerprints, the higher the accumulated match score.

The algorithm is the primary (currently the only) matcher used on the VTB. All results reported in this report were generated using the Bozorth98 matcher. The Bozorth98 matcher is currently being tested and will be compared against current commercial AFIS algorithms in a future report.

2.2.4 Scoring Software

Fingerprint technology evaluations not only require the types of system software components described previously (image decoders, minutiae detectors and matchers, etc.), but they also require a significant investment in the development of scoring and analysis tools. Fortunately, the NIST Image Group has a long history of work in biometric evaluations to draw upon (e.g. [31]).

The scoring software on the VTB is based on a framework of terminology and methods defined in the NIST Human-ID Evaluation Framework (HEF) [33]. HEF was designed to be a general framework used to evaluate any biometric, or combination of biometrics. Recently, a suite of analysis tools based on HEF were developed and used to support the Face Recognition Verification Test (FRVT) 2002 [32]. These same tools have been applied to fingerprints on the VTB.

Using this framework and these tools, three general performance analyses are computed on the VTB. Each of these is described in greater detail in Section 4. They include a simple Receiver Operator Characteristic (ROC) curve described in Section 4.2; a more sophisticated Multi-Trial ROC curve described in Section 5.6; and a Correct Identification vs. Gallery Size analysis described in Section 4.2 and Section 5.10.

3. VTB DATA REPOSITORIES

Fingerprint technology evaluations require a vast amount of data, and typically, the more the better. This section documents the different repositories of fingerprints currently being used in experiments on the VTB at NIST. Some of these repositories are available to the public and some are not.

These repositories represent a collection of different types and sources of fingerprints. There are rolled and plain fingerprints from inked tenprint cards. There are also rolled and plain fingerprints captured from live-scan devices. Given this variety, interesting experiments can be conducted to study the effect of searching heterogeneous types of fingerprints, such as searching a repository of rolled fingerprints with live-scanned fingerprints.

It should be noted that repositories labeled “DHS” came from operational data within the former Immigration and Naturalization Service (INS).

3.1 NIST Special Database 14 (SD14)

NIST Special Database 14 (SD14)	
Description FBI Criminal file – a natural distribution of pattern classes	
Number of Subjects 2700	Instances per Subject 2 fingerprint cards per person
Impression Type Majority Inked Rolled	Finger Positions Captured 10 finger positions segmented from rolled impressions on 10-print card
Capture Device(s) Unknown camera	Availability Public
Data Preparation Segmentation of rolled impressions from the 10-print card was predetermined prior to receipt by NIST	

3.2 NIST Special Database 24 (SD24)

NIST Special Database 24 (SD24)	
Description Contains MPEG-2 (Moving Picture Experts Group) compressed digital video of live-scan fingerprint data	
Number of Subjects 20	Instances per Subject One 10-second video sequence per finger
Impression Type Live-scan Plain	Finger Positions Captured All ten finger positions used in study
Capture Device(s) DFR-90	Availability Public (Note: only five fingers per person on CD)
Data Preparation 4 frames from live-scan video sequences of 4 different finger orientations	

3.3 NIST Special Database 29 (SD29)

NIST Special Database 29 (SD29)	
Description FBI Deceased Criminal File	
Number of Subjects 216	Instances per Subject 2 fingerprint cards per person
Impression Type Inked Rolled & Plain	Finger Positions Captured 10 finger positions segmented from rolled impressions on 10-print card, and 10 additional finger positions segmented from four-finger plain impressions on same 10-print card
Capture Device(s) UMAX PowerLook III flatbed scanner	Availability Public
Data Preparation Segmentation failures of the four-finger plain impressions were manually inspected and corrected to enable maximum yield	

3.4 Immigration and Naturalization Service Recidivist Database

- Fingerprints from the DHS IDENT (Automatic Biometric Fingerprint Identification) System

3.4.1 DHS 2-Finger Images (DHS2)

DHS 2-Finger Images (DHS2)	
Description DHS recidivist cases, the majority of which are border crossing cases with Mexico Environment: border patrol field operations	
Number of Subjects ~600,000 (of ~632,000)	Instances per Subject Minimum of 2 cases per person, where each case contains one right index impression and one left index impression.
Impression Type Live-scan Plain	Finger Positions Captured Right and left index fingers
Capture Device(s) DFR-90	Availability Government use only
Data Preparation Include clean up steps – e.g. mate validation and mismatch detection. (See Section 3.4.1.1)	

3.4.1.1 Matcher-Based Quality Control

The following steps were taken to check for clerical errors in the fingerprint sets. This includes cases where the left index finger is swapped with the right index finger and cases where the same finger was captured twice.

- Given all (right, left) pairs of fingerprints for a person
- Remove finger substitutions from 2-pair cases
 - If and only if 2 (right, left) index finger pairs
 - Match first pair's right finger to both left fingers
 - Match second pair's right finger to both left fingers
 - Look for high-scoring matches
 - Remove likely finger substitutions
- Remove right finger substitutions
 - Find sufficiently "good" left finger image
 - Match all left finger images to each other and select image with sufficiently high score
 - Match all right finger images with "good" left image
 - Look for high-scoring matches
 - Remove likely finger substitutions
- Remove left finger substitutions
 - Find sufficiently "good" right finger image
 - Match all right finger images to each other and select image with sufficiently high score
 - Match all left finger images with "good" right image
 - Look for high-scoring matches
 - Remove likely finger substitutions

3.4.2 DHS 10-Finger Images (DHS10)

DHS 10-Finger Images (DHS10)	
Description DHS Criminal database	
Number of Subjects ~52,000 (of 100,000) 46,000 Background (rolled impressions where four-finger plain segmentation failed)	Instances per Subject One 10-print card per person
Impression Type Live-scan printed into 10-print card Rolled & Plain	Finger Positions Captured 10 finger positions segmented from rolled impressions on 10-print card, and 10 additional finger positions segmented from four-finger plain impressions on same 10-print card
Capture Device(s) Unknown	Availability Government use only
Data Preparation Segmentation of rolled impressions from the 10-print card was predetermined prior to receipt by NIST For plain impressions, only successful automatic segmentation results were used. No manual correction of segmentation results was performed, so a very small number of missegmented results may be included. Automatic segmentation resulted in approximately a 50% yield across all available 10-print cards. All cards used in VTB studies were first consolidated. Consolidations were conducted by inter-matching all cards and applying thresholds. (See Section 5.5.1)	

3.4.3 DHS Consolidation Set of 10-Finger Images (DHS10-C)

DHS Consolidation Set of 10-Finger Images (DHS10-C)	
Description Consolidation set derived from the DHS Criminal database	
Number of Subjects 1021	Instances per Subject 2 fingerprint cards per person
Impression Type Live-scan printed into 10-print card Rolled & Plain	Finger Positions Captured 10 finger positions segmented from rolled impressions on 10-print card, and 10 additional finger positions segmented from four-finger plain impressions on same 10-print card
Capture Device(s) Unknown	Availability Government use only
Data Preparation This set of paired 10-print cards is the byproduct of conducting consolidations on the DHS10 repository. Those cards determined to be consolidations (different instances of the same person's fingers) were removed from DHS10 and set aside for independent study.	

3.5 Department of State Mexican Visa Database (DOS)

Department of State Mexican Visa Database (DOS)	
Description DOS Mexican Visa cases Environment: Mexican Consulates offices	
Number of Subjects ~274,000 (of 288,000) ~6 million Background	Instances per Subject Minimum of 2 cases per person, where each case contains one right index impression and one left index impression.
Impression Type Live-scan Plain	Finger Positions Captured Right and left index fingers
Capture Device(s) DFR-90	Availability Government use only
Data Preparation Include clean up steps – e.g. mate validation and mismatch detection. (See Section 3.4.1.1)	

3.6 Texas Department of Public Safety Database (TXDPS)

Texas Department of Public Safety Database (TXDPS)	
Description Texas DPS records	
Number of Subjects ~225,000 (of 550,000) ~225,000 Background (rolled impressions where four-finger plain segmentation failed)	Instances per Subject 1 fingerprint card per person
Impression Type Majority Inked Rolled & Plain	Finger Positions Captured 10 finger positions segmented from rolled impressions on 10-print card, and 10 additional finger positions segmented from four-finger plain impressions on same 10-print card
Capture Device(s) DBA Image Clear, Model # 5011031	Availability Government use only
Data Preparation Segmentation of rolled impressions from the 10-print card was predetermined prior to receipt by NIST For plain impressions, only successful automatic segmentation results were used. No manual correction of segmentation results was performed, so a very small number of missegmented results may be included. Automatic segmentation resulted in approximately a 50% yield across all available 10-print cards. All cards used in VTB studies were first consolidated. Consolidations were conducted by inter-matching all cards and applying thresholds.	

4. EVALUATION FRAMEWORK

The experiments conducted on the VTB are based on the NIST Human-ID Evaluation Framework (HEF) [33], and the results reported in this section follow a protocol similar to that used in Face Recognition Vendor Test (FRVT) 2002 [32]. The elemental requirement of this framework is that a fingerprint system reports a similarity score when two fingerprint impressions are matched to each other. In general, the higher the score, the more likely the two impressions come from the same finger. Experiments are structured around sets of fingerprint images, and for the VTB, sets of minutiae extracted from these images.

4.1 Terminology and Definitions

An experiment is typically comprised of two general sets of fingerprints. There is the *query* set, a collection of fingerprints whose identities are *unknown* at the time of testing, and there is the *target* set, a collection of fingerprints whose identities are *known*. In more common fingerprint terminology, the query set is the search set to be searched *with*, and the target set is the file set to be searched *on*. Various applications can be represented by matching unknown fingerprints from the query set to known fingerprints in the target set. As comparisons are computed, matcher scores are stored in a *similarity matrix* where the ij -th element in the matrix corresponds to the similarity between the i -th fingerprint of the target set compared to the j -th fingerprint of the query set.

Once a similarity matrix is populated with matcher scores, performance statistics are computed. If subsets of the query and target fingerprints are known to share a common trait, then performance statistics may be computed on just these subsets in order to isolate and study the effect of these traits. The subset used from the query set is referred to as the *probe* set, and the subset used from the target set is referred to as the *gallery* set.

The scores in the similarity matrix fall into two general categories. A score computed between a probe and gallery belonging to the *same* person is referred to as a *match*, while a score computed between a probe and gallery belonging to *different* persons is referred to as a *non-match*. (Note that the terms, match and non-match, are being used here to characterize whether the probe and gallery fingerprints are from the same person, and not whether the matcher actually achieved a correct identification, which is also often referred to as a correct match or hit.) Significant insights into the performance of a fingerprint system may be gained by analyzing and comparing the distribution of match scores to the distribution of non-match scores.

4.2 Verification vs. Identification

To understand what is involved to develop biometric standards and to conduct biometric technology evaluations, it is helpful to know that biometric applications are typically categorized into two general types: verification and identification.

The term *verification* is used to describe the process of confirming that a person is who he/she claims to be by matching their biometric record against that of their claimed identity. It is a one-to-one comparison. *Identification* is a term used to describe the process of matching a biometric record from a single unknown person against an entire repository of similar biometric records in

order to determine the identity of the owner of the biometric record. It is a one-to-many comparison.

The purpose of a verification system is to simultaneously perform two tasks. The first is to correctly verify the identity of a person when the claim is legitimate. The second is to reject people who are not who they claim to be. Unfortunately, there is a trade-off between these two tasks, and one cannot simultaneously maximize the performance of both tasks.

The performance statistic for verifying the identity is the probability of correct verification or *true accept rate (TAR)*. This is the probability that a system will verify the identity of a legitimate claim. The performance statistic for rejecting false claims is referred to as the *false accept rate (FAR)*. This is the probability that a false claim will be accepted as being true; i.e., someone fools the system and an unauthorized person is granted access.

A Receiver Operator Characteristic (ROC) analysis measures the trade-off between the true accept rate and the false accept rate. The result is a curve which serves as a primary measurement of verification performance. For the purposes of discussion and comparison, two points of interest are cited from the ROC curves presented in this paper. The first is the true accept rate achieved at a false accept rate of 1%. The second is the false accept rate achieved at a true accept rate of 98%. (These numbers, although somewhat arbitrary, are representative of an acceptable operating range for many applications.)

Identification performance is measured by determining the ability of a biometric system to identify an individual in a large database, given a single unknown biometric record. In the process used in FRVT, the probability of correct identification at rank one (the system's top choice) is computed. As the size of the database used for identification increases, the probability of an incorrect match having higher score than the correct match increases, and the rate of correct matches at rank-1 decreases. Therefore a curve is generated, depicting the effect of database size on the probability of correct identification. This serves as a useful measurement of identification performance.

The FRVT nomenclature defines a watch list as an application in which probes are detected and identified using a combination of score and rank-based operating thresholds. It should be noted that, to date, no fingerprint-based watch list applications have been studied on the VTB.

5. STUDIES AND RESULTS

5.1 Overview of Studies

5.1.1 Small-Scale Studies

Section 5.2, Inked, Rolled Impression Verification Study with SD14, reports on the results of early studies measuring the verification results of a small set of rolled fingerprints. These results are intended to serve as a best-case baseline to be used in comparisons with other studies.

Section 5.3, Live-Scan, Plain Impression Verification Study with SD24, reports the verification results of live-scan plain fingerprints from a small data set.

Section 5.4, Inked, Rolled vs. Plain Impression Verification Study with SD29, reports the verification results of a small set of fingerprints from inked cards in which rolled-to-rolled, rolled-to-plain (four-finger segmented), and plain-to-plain comparisons were made for all fingers. These results provide an important baseline for performance of rolled and plain fingerprints, and for finger-by-finger comparisons.

Section 5.5, Live-Scan, Rolled vs. Plain Impression Verification Study with DHS10-C, reports the verification results of a larger set of fingerprints from cards printed with live-scanned fingerprints in which rolled-to-rolled, rolled-to-plain (four-finger segmented), and plain-to-plain comparisons were made for right thumbs and right index fingers.

5.1.2 Large Scale Studies

Section 5.6, Large Scale Live-Scan Verification Study with DHS2, reports the verification results of a large set of low-quality operational live-scan fingerprints from the DHS. These results are used to demonstrate the statistical effect of sample size when measuring matcher performance.

Section 5.7, Large Scale Live-Scan Verification Study with DOS, reports the verification results of a large set of operational live-scan fingerprints from the Department of State.

Section 5.8, Large Scale Inked Verification Study with DHS10, reports the verification results of plain-to-rolled matching of right thumbs and right index fingers, taken from a large set of inked fingerprint cards from the DHS.

Section 5.9, Large Scale Inked Verification Study with TXDPS, reports the verification results of plain-to-rolled matching of right thumbs and right index fingers, taken from a large set of inked fingerprint cards from the Texas Department of Public Safety.

Section 5.10, Large Scale Identification Study with DHS2, reports the identification results from the DHS2 live-scan data, measuring performance in terms of correct identifications at rank-1. These results also show the effect of gallery size on rank-based identification.

Section 5.11, Large Scale Identification Study with DOS, reports the identification results from the DOS live-scan data, measuring performance in terms of correct identifications at rank-1. These results also show the effect of gallery size on rank-based identification.

5.1.3 Other Studies

Section 5.12.1, Score-Based Fusion Using SD29, explores how the index finger and thumb results from the SD29 study can be fused at the score level to improve matcher performance.

Section 5.12.2, Score-Based Fusion Using DHS10-C, explores how the index finger and thumb results from the SD29 study can be fused at the score level to improve matcher performance.

Section 5.12.3, Rank and Score-Based Fusion Using DHS2, explores how the index finger results from the SD29 study can be fused at the rank level to improve matcher performance.

Section 5.13, Person Variation Study with DHS2, explores whether some people's fingerprints are intrinsically more difficult to match than others.

5.2 Inked, Rolled Impression Verification Study with SD14

Early experiments on the VTB were designed to derive a baseline of performance on traditional rolled fingerprints from FBI tenprint cards. This served to validate the software running on the VTB, while at the same time it provided a level of performance against which subsequent studies involving plain and rolled impressions could be compared.

A verification study was constructed with images of inked, rolled fingerprints from NIST Special Database 14 (SD14). SD14 contains 2700 mated pairs of FBI tenprint cards from 2700 different people. Using the rolled impressions from these cards, probe fingerprints were selected from the search (query set) cards in the repository, while gallery fingerprints were selected from the file (target set) cards in the repository.

A verification study was conducted whereby performance of left and right thumb and index fingers were compared. Figure 3 plots ROC curves resulting from four different 2700×2700 similarity matrices. Note that in general, index fingers performed better than thumbs, and that right fingers performed better than left.

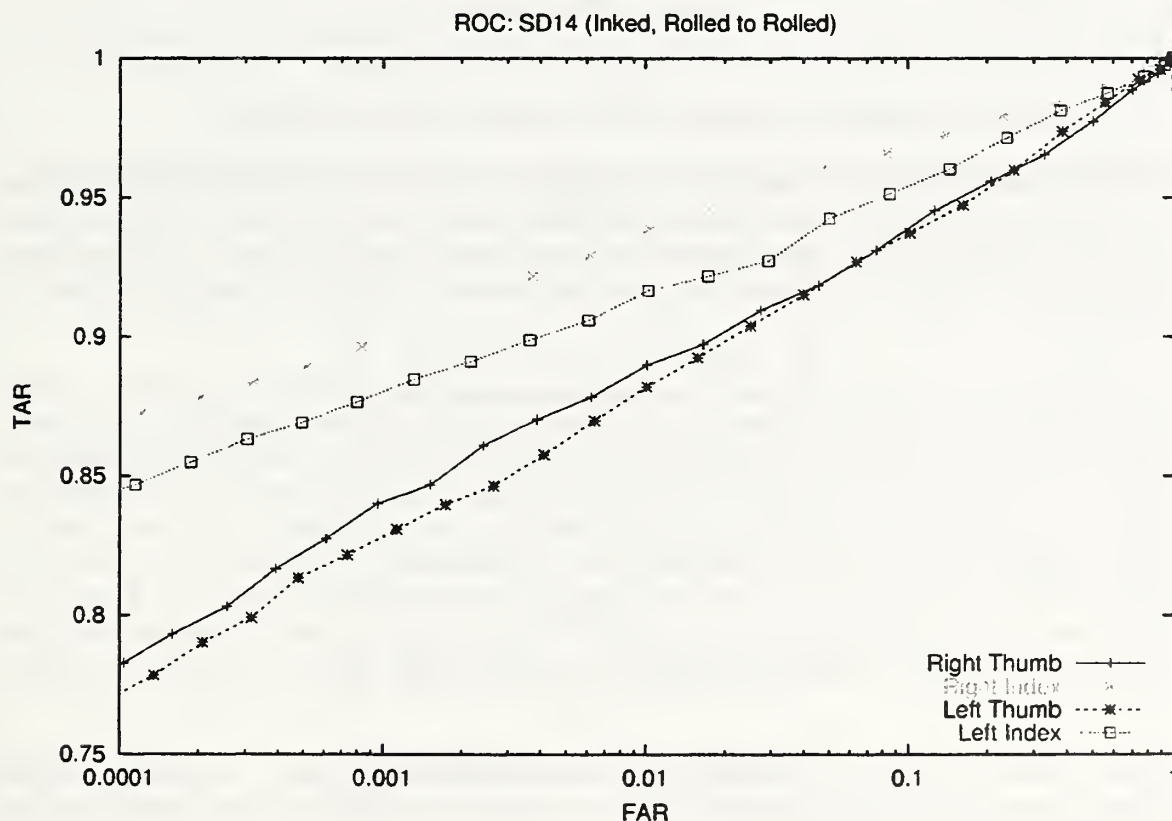


Figure 3. SD14 Verification Study – Comparison of right and left thumbs and index fingers

SD14	
FAR @ 1% & TAR @ 98%	
Right Thumb	
(1%, 89%)	(51%, 98%)
Right Index	
(1%, 94%)	(23%, 98%)
Left Thumb	
(1%, 88%)	(55%, 98%)
Left Index	
(1%, 92%)	(38%, 98%)

Table 1. SD14 Results

5.3 Live-Scan, Plain Impression Verification Study with SD24

A technical issue facing next generation border control systems is what level of performance can be expected when searching legacy data comprised of inked, rolled impressions with new plain impressions captured with a live-scan device. To begin exploring these issues, NIST collected and published a sample of live-scan fingerprint impressions called NIST Special Database 24 (SD24).

A verification study was constructed with live-scan, plain fingerprints from SD24. SD24 contains 20 people, each contributing a 10 second video sequence per finger. This study compared the performance between all five fingers: thumbs, index, middle, ring, and little fingers. Four frames of video were selected per finger, and left and right corresponding finger positions were combined. In all, five similarity matrices were computed, each of size 160×160. These dimensions are the result of (160 fingerprints = 20 people × 4 impressions × 2 left & right hands). Due to the limited size of SD24, the same 160 impressions were used for both the probe set and the gallery set (ignoring the comparisons of each image to itself). Note that standard ROC protocol calls for a separate set of impressions to be used between the probe and gallery sets.

Figure 4 plots ROC curves resulting from the five different fingers. In general, thumbs performed best; next, index and middle fingers performed comparably to each other; then ring fingers; followed lastly by little fingers.

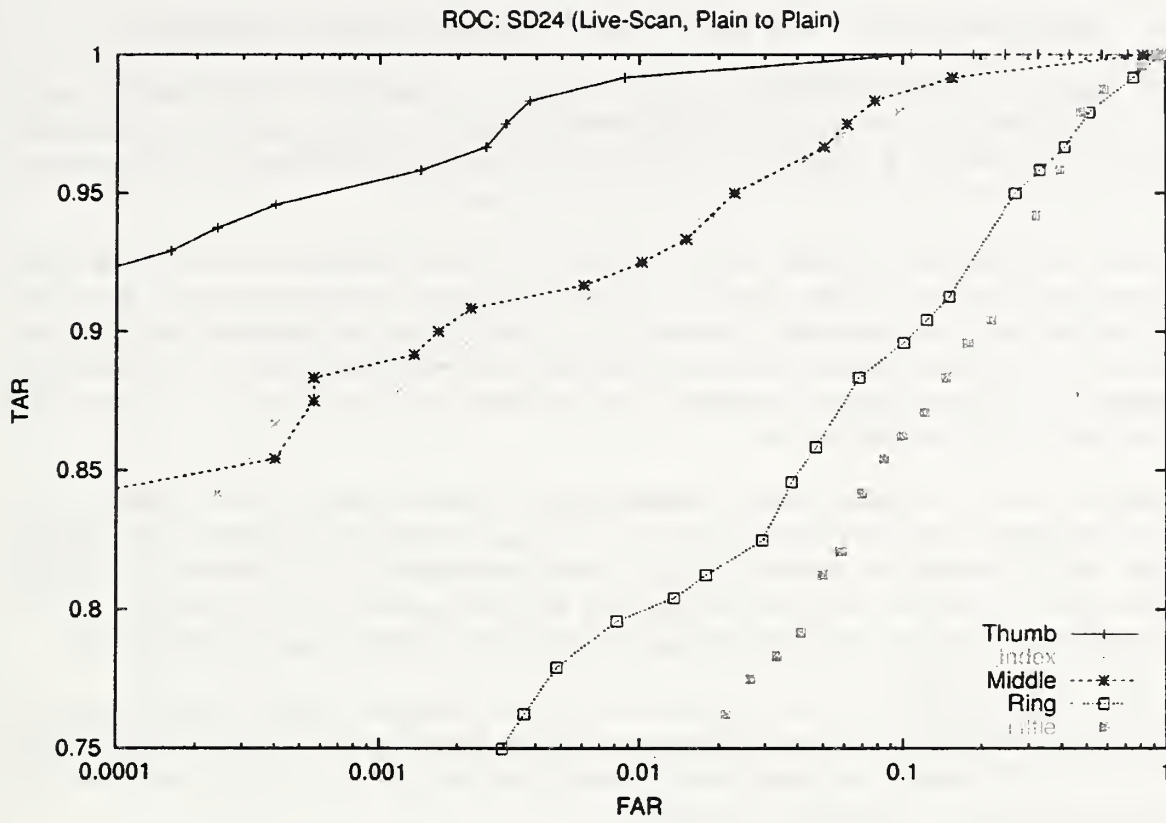


Figure 4. SD24 Verification Study – Comparison of thumb, index, middle, ring, and little fingers

SD24	
FAR @ 1% & TAR @ 98%	
Thumb	
(1%, 99%)	(0.4%, 98%)
Index	
(1%, 94%)	(10%, 98%)
Middle	
(1%, 92%)	(8%, 98%)
Ring	
(1%, 80%)	(51%, 98%)
Little	
(1%, 70%)	(48%, 98%)

Table 2. SD24 Results

5.4 Inked, Rolled vs. Plain Impression Verification Study with SD29

NIST Special Database 29 (SD29) is a small collection of mated pairs of FBI tenprint cards. There are 216 people in this repository, each contributing two *complete* cards. Although small, SD29 is significant as it contains *all* impressions on the card, including plain impressions. (SD14 has many more people, but only contains rolled impressions.)

As seen in Figure 2, two boxes are provided on a standard tenprint card in which plain impressions of a person's index, middle, ring, and little fingers are entered together from their right and left hands. As mentioned in Section 2.2.2, a four-finger plain segmenter was developed and run on the VTB to separate the four in each box into individual images. As a result, two complete sets of fingerprints were extracted from each card in the repository, a rolled set of ten fingers and a corresponding plain set.

A series of verification studies were conducted on the rolled and plain impressions from SD29. Probe and gallery sets were selected so as to compare performance between three different modes: rolled impressions searched against rolled impressions, plain impressions searched against plain impressions, and plain impressions searched against rolled impressions. These three modes were evaluated for all five fingers: thumb, index, middle, ring, and little fingers.

For each mode, a similarity matrix of size 432×432 was computed. These dimensions are the result of 216 people with left and right corresponding finger positions (×2) combined in the matrix. The 432 probe fingerprints were selected from card set 'a' in SD29, while the 432 gallery fingerprints were selected from card set 'b'.

The results of these experiments are shown in the five following figures. For example, Figure 5 plots three ROC curves (one for each search mode) derived from thumb impressions. The top curve in the figure corresponds to the performance of searching rolled impressions against rolled impressions; the middle curve corresponds to the performance of searching plain impressions against plain impressions; and the bottom curve corresponds to the performance of searching plain impressions against rolled impressions.

In general, thumbs performed best; next, middle fingers, then index fingers; then ring fingers; followed lastly by little fingers. It should also be noted that rolled-to-rolled searches performed consistently and significantly better, while plain-to-plain and plain-to-rolled were closer together with plain-to-plain performing frequently better.

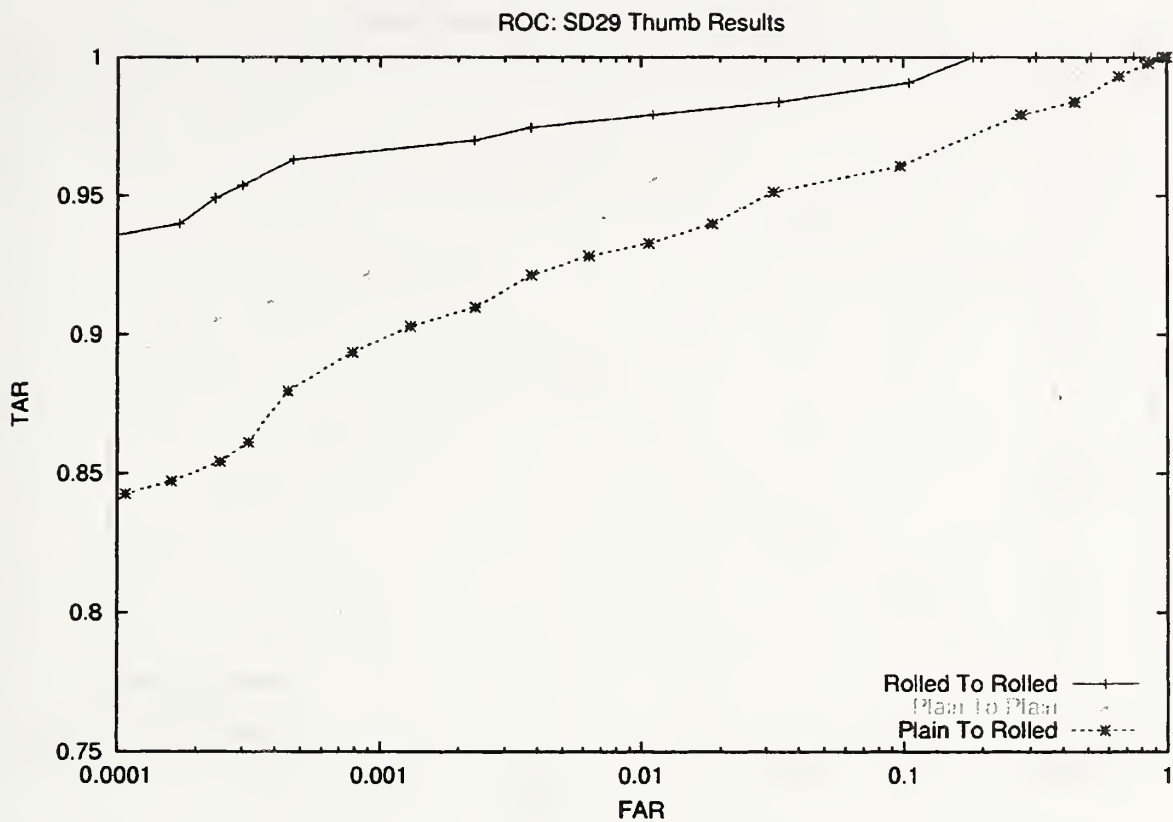


Figure 5. SD29 Thumb Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

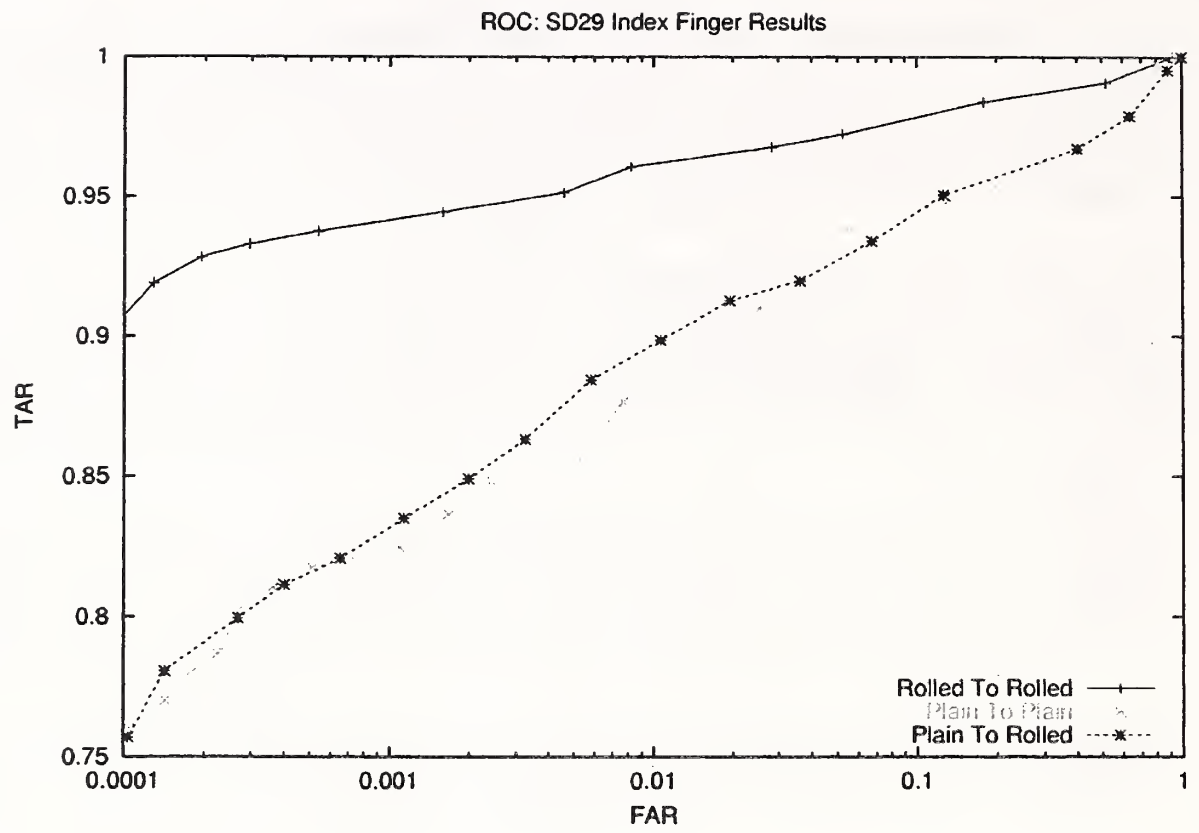


Figure 6. SD29 Index Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

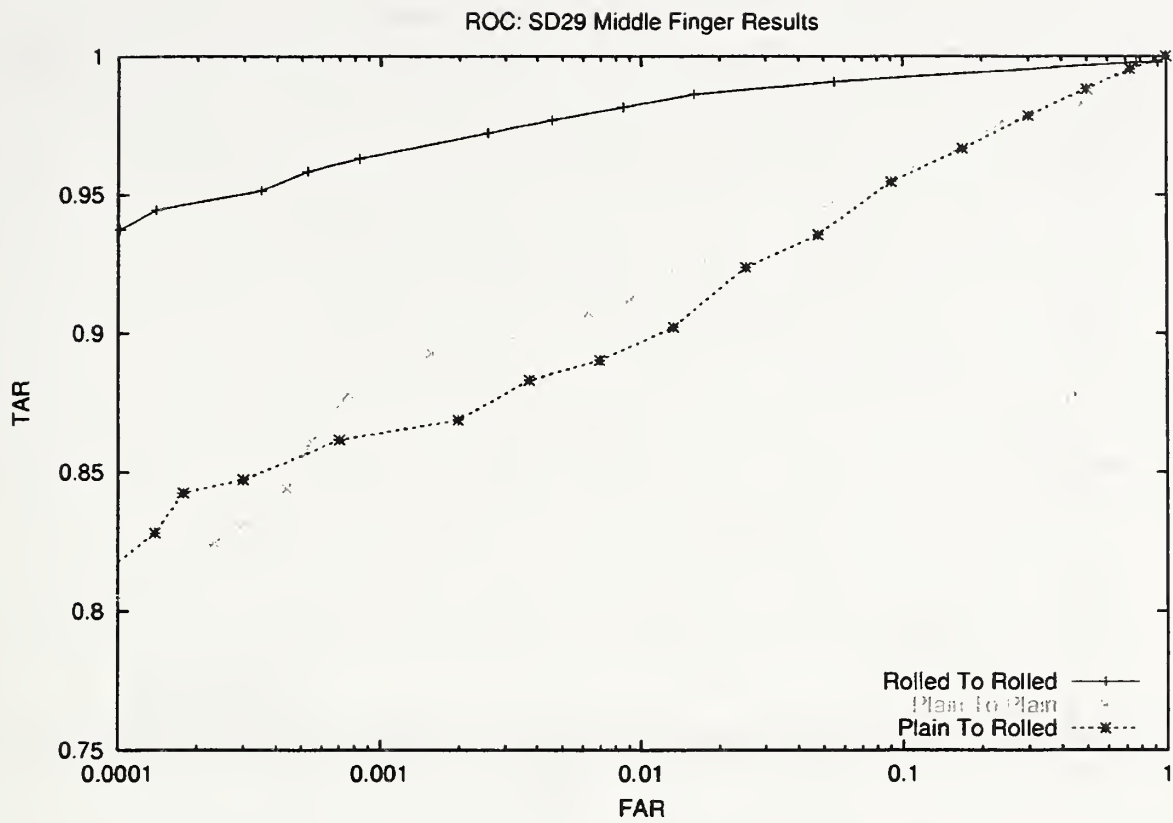


Figure 7. SD29 Middle Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

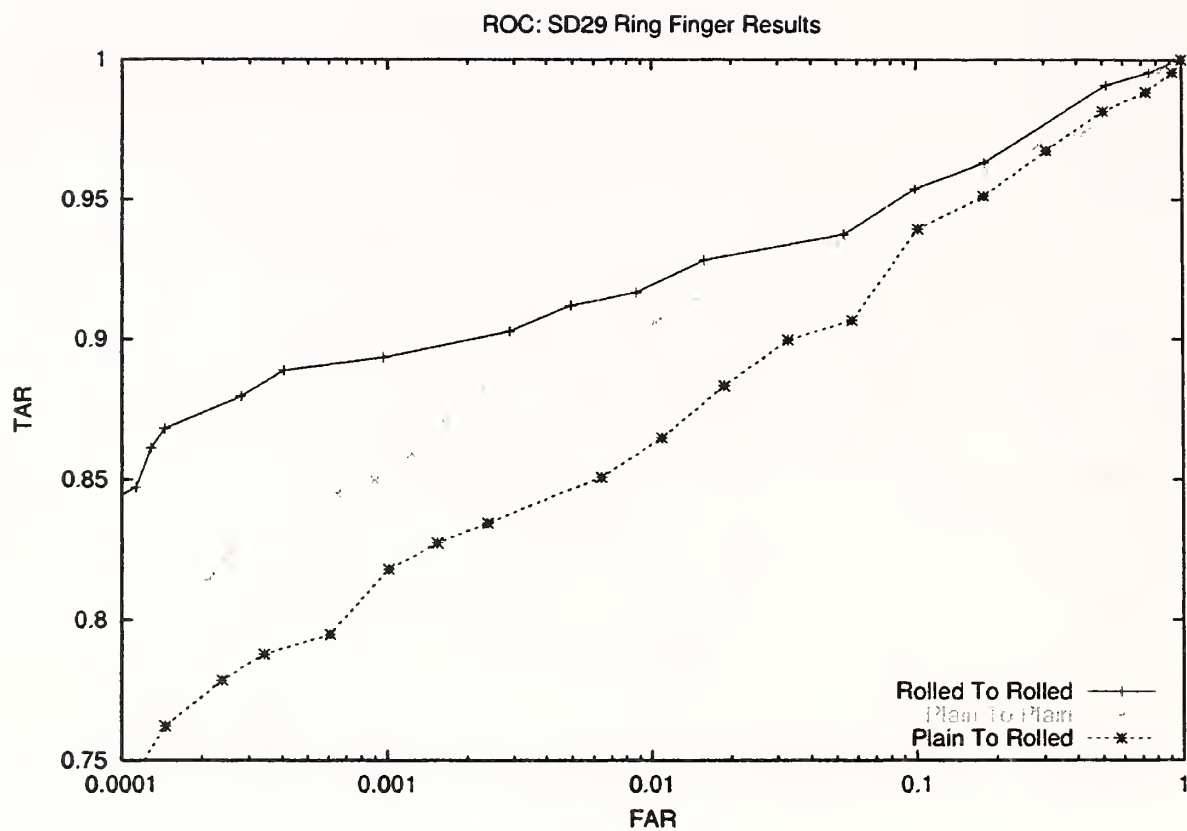


Figure 8. SD29 Ring Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

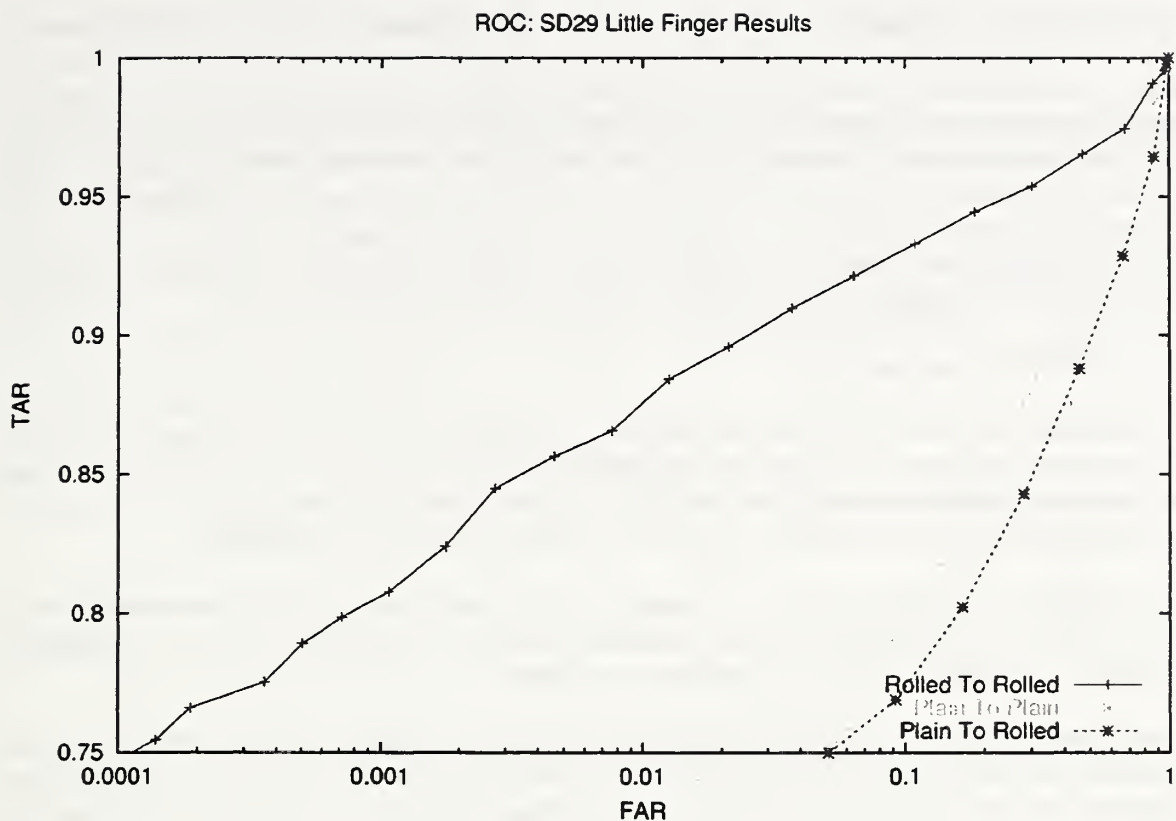


Figure 9. SD29 Little Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

SD29						
FAR @ 1% & TAR @ 98%						
	Rolled-to-Rolled		Plain-to-Plain		Plain-to-Rolled	
Thumb	(1%, 98%)	(1%, 98%)	(1%, 95%)	(19%, 98%)	(1%, 93%)	(28%, 98%)
Index	(1%, 96%)	(18%, 98%)	(1%, 88%)	(43%, 98%)	(1%, 90%)	(64%, 98%)
Middle	(1%, 98%)	(1%, 98%)	(1%, 91%)	(48%, 98%)	(1%, 90%)	(30%, 98%)
Ring	(1%, 95%)	(40%, 98%)	(1%, 91%)	(62%, 98%)	(1%, 86%)	(50%, 98%)
Little	(1%, 88%)	(70%, 98%)	(1%, 70%)	(90%, 98%)	(1%, 66%)	(93%, 98%)

Table 3. SD29 Results

5.5 Live-Scan, Rolled vs. Plain Impression Verification Study with DHS10-C

Repositories containing mated pairs of cards, such as SD29, are rare to come by. AFIS systems typically archive only one card per person. Search cards are submitted for search, but they are not permanently stored. While somewhat rare, repositories with mated pairs of cards are very useful for technology evaluations. They provide the ability to compare three fundamental modes of searching that current border entry systems must consider and manage: rolled-to-rolled, plain-to-plain, and plain-to-rolled. This section describes a study conducted on a larger collection of mated card pairs, referred to as DHS10-C.

5.5.1 DHS10 Consolidation

As described in Section 3.4.2, NIST has acquired a collection of tenprint card images from DHS. The cards are individual records from the agency's criminal database. Unknown to NIST was whether each fingerprint card in the collection uniquely represented one person, or did multiple cards belonging to the same person exist? This is what is known as *consolidation*.

Consolidation is an important topic to biometric system performance, both operationally and in the laboratory. Unknowingly having redundant records in a repository can cause unexpected confusions, overhead, and skewed performance statistics.

Simply put, imagine an experiment where a file repository is seeded with (has added to it) a set of fingerprint card images belonging to a certain person. Images of a second set of impressions from the same person are then used to search the seeded repository. If the system determines the search impressions sufficiently match the person's file impressions, then the system has performed a successful identification. But what if the repository, prior to seeding, unknowingly contains another set of impressions belonging to the person? Now, when the seeded system is searched, there will be confusion, and if the system determines the unknown file card (rather than the seeded card) to be the rank-1 match, then it may be incorrectly concluded that the system failed to make the proper identification. (Note that this example is based on simple rank-1 performance, and that AFIS systems typically operate on more robust measurements.)

An automated process was developed to detect and remove consolidation cases from the DHS10 repository. Originally, 100K tenprint card cases were provided by DHS; of these, ~54K were automatically segmented into separate plain impressions. To detect consolidation cases, the rolled impressions from the set of ~54K were fully searched against themselves, and thresholds were applied to determine consolidations. Of the ~54K, approximately 2K cases were removed.

The consolidation cases that were removed inherently represent a collection of mated cards. As a result, 1021 people were determined to have card mates, and these card images were gathered into their own repository, called DHS10-C.

5.5.2 DHS10-C Results

A study, similar to the one conducted with SD29, was conducted using the somewhat larger set of mated rolled and plain impressions of DHS10-C. Probe and gallery sets were selected to compare the performance of rolled-to-rolled, plain-to-plain, and plain-to-rolled searches. These

three modes were evaluated using both thumbs and index fingers. The results reported here are for right fingers. Left finger results were also computed and produced similar results.

A similarity matrix of size 1021×1021 was computed for each of the three search modes. Images from the first card mate were used as the gallery set, while images from the second card mate were used as the probe set.

Figure 10 plots three ROC curves (one for each search mode) derived from right thumb impressions. The top curve in the figure corresponds to the performance of searching plain impressions against plain impressions. Just below is the rolled-to-rolled curve. This is quite different from the results achieved on thumbs with SD29 results in Figure 5. The lowest performing mode is plain-to-rolled.

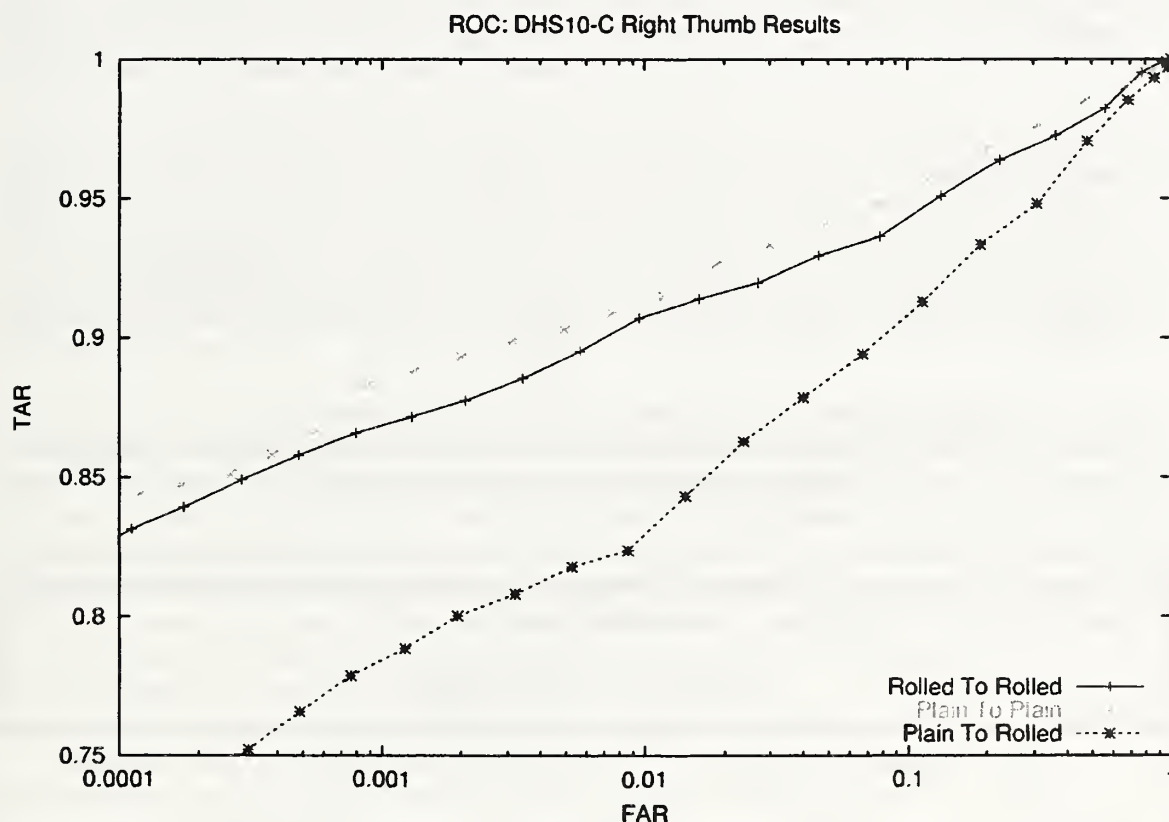


Figure 10. DHS10-C Right Thumb Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

Figure 11 shows the ROC results from DHS10-C right index fingers. Notice that overall performance is much lower with index fingers than with thumbs. Also, the rolled-to-rolled mode performs significantly higher than the other two modes (which is more consistent with SD29 results), and there is significant separation between plain-to-plain and plain-to-rolled modes.

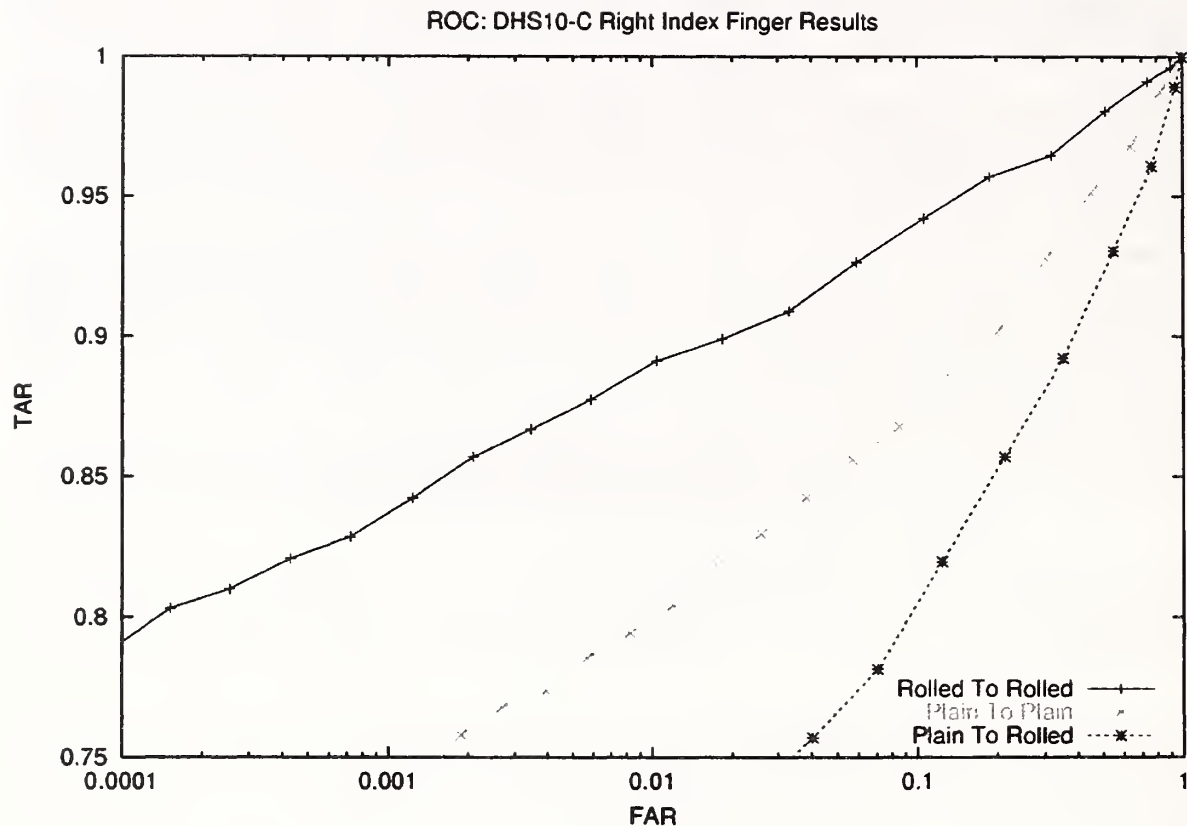


Figure 11. DHS10-C Right Index Finger Verification Study – Comparison of rolled-to-rolled, plain-to-plain, and plain-to-rolled

DHS10-C						
FAR @ 1% & TAR @ 98%						
	Rolled-to-Rolled		Plain-to-Plain		Plain-to-Rolled	
Right Thumb	(1%, 91%)	(56%, 98%)	(1%, 91%)	(40%, 98%)	(1%, 83%)	(62%, 98%)
Right Index	(1%, 89%)	(51%, 98%)	(1%, 80%)	(75%, 98%)	(1%, 71%)	(87%, 98%)

Table 4. DHS10-C Results

5.6 Large Scale Live-Scan Verification Study with DHS2

Up to this point, the VTB experiments documented in this report have been quite constrained by the size of the repositories used. The DHS2 repository contains nearly 600K people and provides the opportunity to conduct significantly larger experiments. In fact, when processing repositories of this size, the major constraint shifts from the amount of data to the number of cycles available to compute on the VTB.

An important question to explore is, “What amount of data is needed to get statistically reliable results from performance evaluations?” The most significant factor to be considered is the characteristic quality of the fingerprints in the repository. If one computes performance statistics on an overly small sample of fingerprints, results will be quite unreliable. This instability is observed as significant variation in performance metrics when subsequent independent samples of the same size are computed and compared. As the size of the sample increases, the variation observed between independent trials will become more stable.

A verification study using DHS2 was designed to explore this issue more closely. Pairs of mated right index finger impressions were compared. The first impression in the pair was used as the probe image, and the second was used as the gallery image. To study the amount of variation in computed performance, a random set of 60K people were selected from the repository. This list was then subdivided into ten independent sets of 6K people, and the corresponding probe and gallery images were compared and resulting matcher scores were compiled into ten 6K×6K similarity matrices.

To look at the variation in performance, one could simply plot and visually compare the ten ROC curves, each corresponding to one of the ten similarity matrices. Figure 12 plots a more sophisticated and useful analysis, called a *Multi-Trial ROC*. The blue curve in the graph plots the *mean* of the ten ROC curves. The small clusters of gray points along the curve contain synchronized values extracted from each of the ten ROC curves. The spread of the points within these clusters represent the variation in performance between each of the ten random trials. The red ellipse overlaying each cluster represents a statistically standardized amount of variance across the trials. The radius of each ellipse is (2×Standard Error), measured from the points in the cluster along both the x-axis and the y-axis.

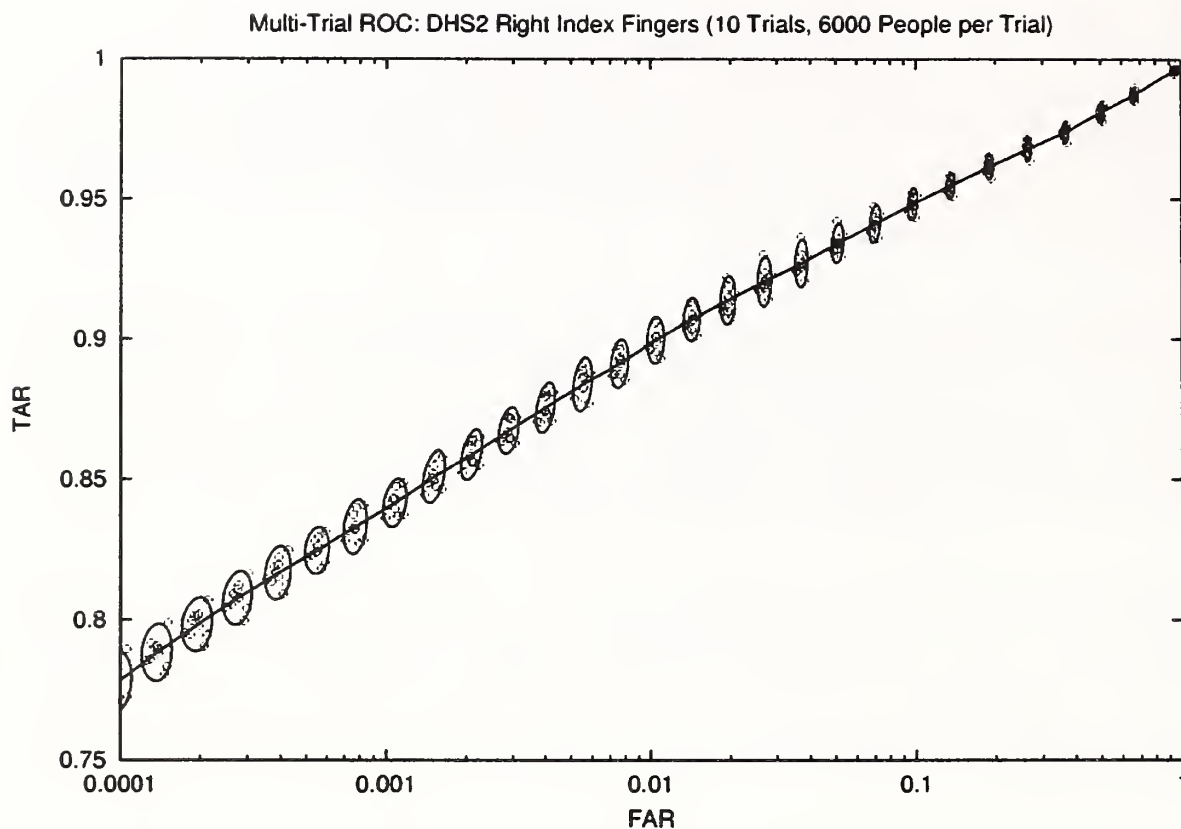


Figure 12. DHS2 Plain-to-Plain Right Index Finger Large Scale Verification –Multi-Trial ROC

DHS2	
FAR @ 1% & TAR @ 98%	
Right Index	
(1%, 90%)	(50%, 98%)

Table 5. DHS2 Results

As one can see in Figure 12, only a moderate amount of fluctuation is observed among the ten ROC curves. An interesting question is, “What might the variation be if the size of the similarity matrices were cut in half?”

This is relatively easy to explore using the ten 6K×6K similarity matrices which have already been computed. There are two complete 3K×3K similarity matrices within each of the 6K×6K

matrices. For example, there is one in the top-left quadrant and the other in the lower-right quadrant. For reasons discussed in Section 6.1, these simple quadrants were not studied here, but rather a random selection of people was used between the first and second 3K×3K sets.

ROC curves were computed from each of the ten 3K×3K similarity matrices in the first set, and a Multi-Trial ROC analysis was conducted. The same analysis was conducted on the ROC curves computed from the ten 3K×3K similarity matrices in the second set. If a sample size of 3K×3K is sufficient, then one would expect a similar amount of variation between the ROC curves computed from the 3K×3K similarity matrices and the 6K×6K similarity matrices. If quality were consistent across people in DHS2, then one would expect similar mean results between the 3K×3K similarity matrices and the 6K×6K similarity matrices.

There are three Multi-Trial ROC curves plotted in Figure 13. The ROC belonging to the “parent” ten 6K×6K similarity matrices is plotted in gray and is obscured by the other two curves belonging to the 3K×3K submatrix results. As can be seen, the two 3K×3K curves significantly overlap with each other and the 6K×6K results. The most notable difference is the ellipses associated with the blue curve (First 3000), which are typically larger than those associated with the other two. From these curves, it is concluded that the 3K×3K submatrix sets reasonably represent the parent 6K×6K set except with slightly greater variance.

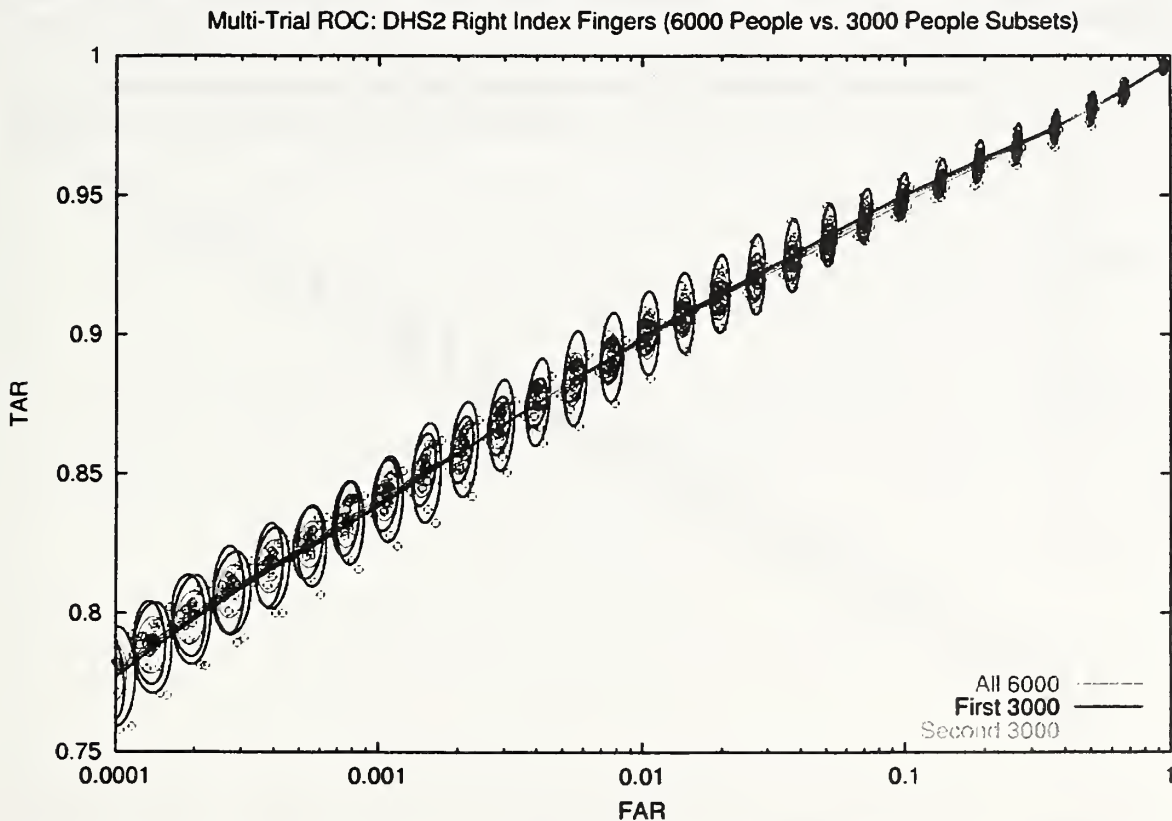


Figure 13. DHS2 Right Index Finger Large Scale Verification –Submatrix variation of Multi-Trial ROC

5.7 Large Scale Live-Scan Verification Study with DOS

NIST has acquired a collection of live-scanned index fingerprints from DOS. This repository has many similarities to that of DHS2; however, the DOS fingerprints were collected by different personnel, at different locales, and under different conditions. It is very interesting to explore how the quality of the DHS2 and DOS repositories may differ.

A verification study using DOS was conducted to assess the quality of this repository. Pairs of mated index finger impressions were compared. To study the amount of variation in computed performances, a Multi-Trial ROC analysis was conducted similar to the DHS2 study in the previous section.

A random set of 30K people were selected from the DOS repository. This list was then subdivided into ten independent sets of 3K people. For each person, a pair of right index fingers *and* a pair of left index fingers were selected and fully compared to all other fingerprints. The first impression in the pair was used as the gallery image, and the second was used as the probe image. The matcher scores were compiled into ten 6K×6K similarity matrices.

Figure 14 plots the resulting Multi-Trial ROC curve. Comparing this graph with Figure 12, DOS performs slightly better at lower false accept rates than DHS2, but their statistical ellipses significantly overlap.

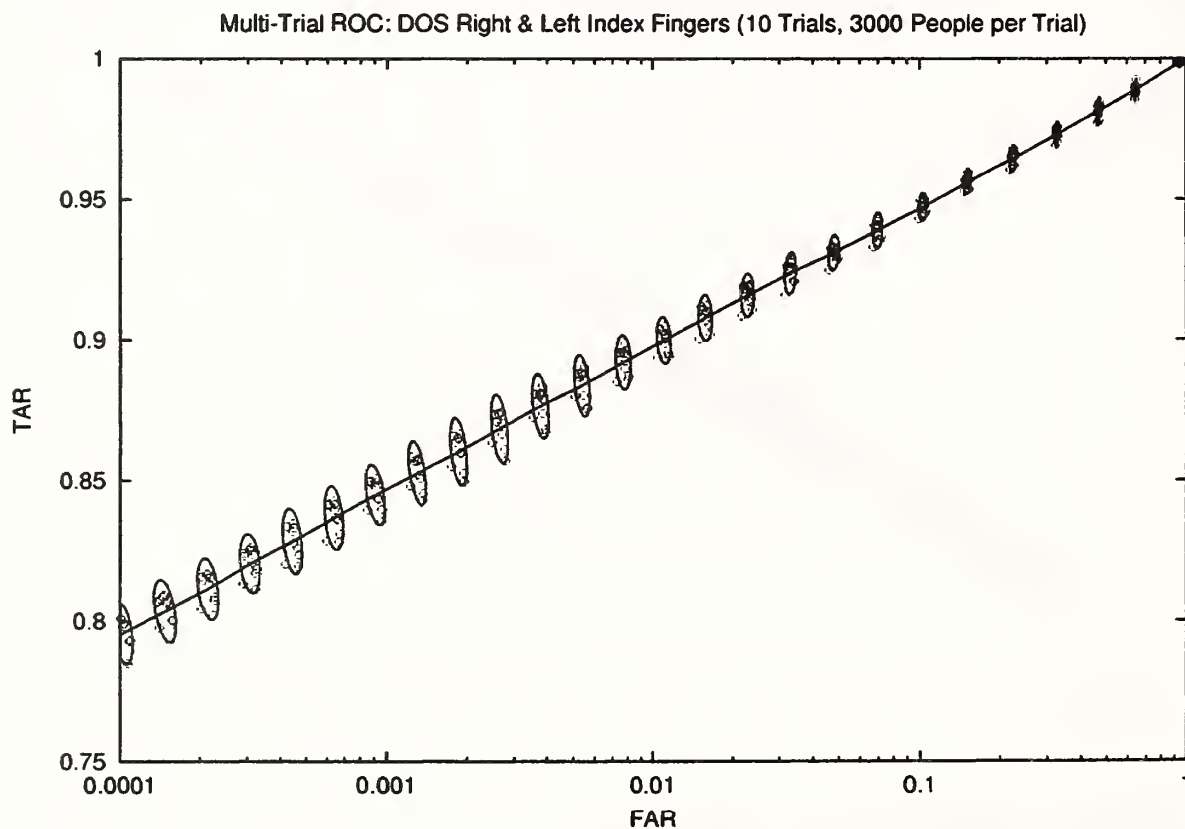


Figure 14. DOS Index Finger Large Scale Verification –Multi-Trial ROC

DOS	
FAR @ 1% & TAR @ 98%	
Right & Left Index	
(1%, 90%)	(47%, 98%)

Table 6. DOS Results

Two sets of 3K×3K similarity matrices were extracted from the ten 6K×6K matrices, the first set from the upper-left quadrant of scores, and second set from the lower-right quadrant of scores. Each quadrant contains right and left index finger comparisons from 1500 people, resulting in a 3K×3K submatrix. A Multi-Trial ROC was computed from both sets of submatrices. The resulting curves are overlaid in Figure 15.

Looking at the three mean ROC curves in the graph, the two 3K×3K curves (First 1500 & Second 1500) are very similar to the 6K×6K curve (All 3000). The most notable difference is the ellipses associated with the green curve (Second 1500) are significantly larger than those associated with the other two. Based on these results, it is concluded that the 3K×3K subsets are reasonably representative of the parent 6K×6K set.

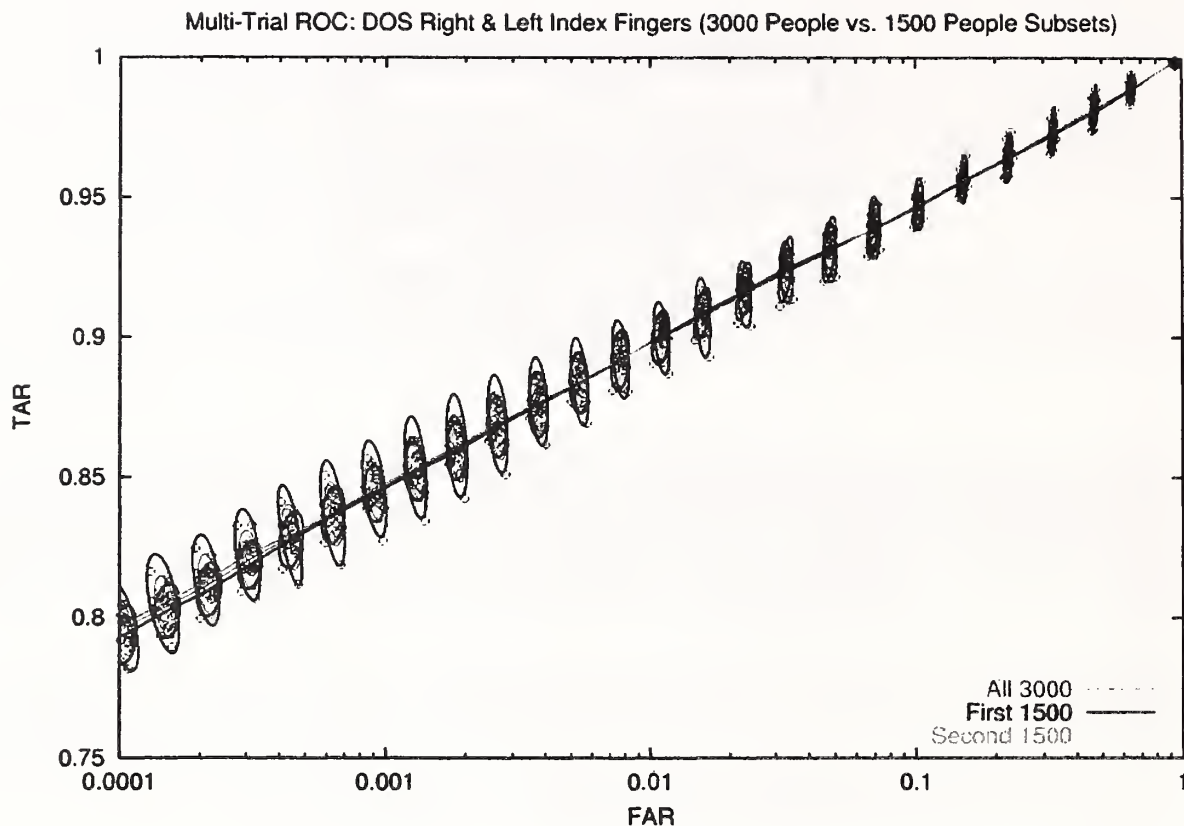


Figure 15. DOS Index Finger Large Scale Verification -- Submatrix variation of Multi-Trial ROC

5.8 Large Scale Inked Verification Study with DHS10

NIST acquired a collection of tenprint card images from DHS in addition to the live-scanned index fingerprints in DHS2. The fingerprint card repository, DHS10, contains two complete sets of finger impressions per card, one set of ten rolled impressions, and one set of ten plain impressions. Each card was provided to NIST as a sequence of 14 images, corresponding to the isolated fingerprint boxes on the card. The NIST four-finger plain segmenter was used to create images of individual plain impressions.

A verification study was designed to evaluate the performance of matching the plain impressions of index fingers in DHS10 to their corresponding rolled impressions. A Multi-Trial ROC analysis was conducted to study the quality in this repository and the variation within results.

A list of 51,440 people (nearly the entire DHS10 repository) was selected. This list was then subdivided into ten independent random sets of 5144 people. For each person, a pair of right thumbs was selected and fully compared to all other thumbs in the set. The rolled impression in each pair was used as a gallery image, and the plain impression was used as a probe image. The matcher scores were compiled into ten 5144×5144 similarity matrices. The same process was then repeated by selecting pairs of (rolled, plain) right index finger impressions.

Figure 16 plots the resulting Multi-Trial ROC curves, one for right thumbs, and the other for right index fingers. Notice that DHS10 performance is lower than the plain-to-plain comparisons of DHS2 right index fingers in Figure 12, and performance is lower than the plain-to-rolled comparisons of SD29 in Figure 5 and Figure 6. This is a clear indication that the data in DHS10 is more difficult to match than the data in these other repositories.

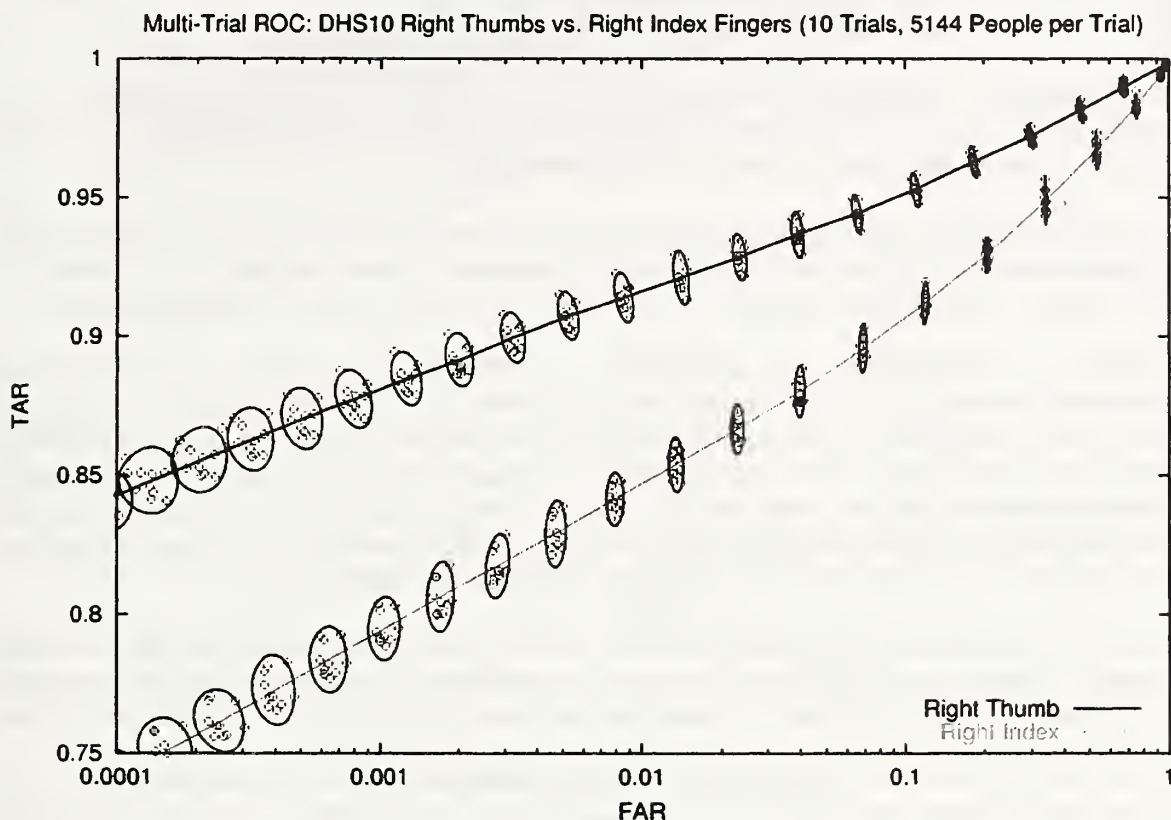


Figure 16. DHS10 Right Thumb vs. Right Index Finger Large Scale Verification –Multi-Trial ROC (Plain-to-Rolled)

DHS10	
FAR @ 1% & TAR @ 98%	
Right Thumb	
(1%, 92%)	(47%, 98%)
Right Index	
(1%, 85%)	(75%, 98%)

Table 7. DHS10 Right Thumb vs. Right Index Results

A study was conducted whereby the two sets of ten 5144×5144 similarity matrices were divided into upper-left and lower-right quadrants, and a Multi-Trial ROC analysis was separately performed on each set of submatrices. The results were similar to the analysis conducted on the DOS repository and plotted in Figure 15. The submatrix curves overlapped closely with the full matrix curve with slightly larger variance ellipses.

5.9 Large Scale Inked Verification Study with TXDPS

A large number of fingerprints from tenprint cards were acquired from the Texas Department of Public Safety. A repository of rolled impressions and segmented plain impressions was created using the same process used to create the DHS10 repository.

A verification study was designed to evaluate the performance of matching the plain impressions of thumbs and index fingers in TXDPS to their corresponding rolled impressions. A Multi-Trial ROC analysis was conducted to study the quality in this repository and the variation in results.

A random set of 30K people were selected from TXDPS. This list was then subdivided into ten independent sets of 3K people. For each person, a (rolled, plain) pair of right thumb impressions and a (rolled, plain) pair of left thumb impressions were selected. The rolled right and left thumb impressions were combined into one set of gallery images. The plain impressions were used as probe images, and they were matched fully with the combined gallery set, whether left or right. The matcher scores were compiled into ten $6K \times 6K$ similarity matrices. The same process was then repeated by selecting (rolled, plain) pairs of right and left index fingers.

Figure 17 plots the resulting Multi-Trial ROC curves, one for thumbs, and the other for index fingers. Notice that TXDPS performance is significantly higher than the plain-to-rolled comparisons of DHS10 in Figure 16; however, the variance ellipses for TXDPS are much larger.

Notice the apparently convex shape of the ellipses in Figure 17. This is due to the size of the ellipses being plotted on a log scale along the x-axis (a *semilog* graph). For TXDPS data, the variance is high enough so that this experiment is “large signal” (visibly non-linear). In the previous curves the ellipses were small enough so that they retained their elliptical shape under a logarithmic transformation due to the small changes in test results meeting a “small signal” criterion. (A sufficiently small change in the independent variable along a continuous function can be approximated using a linear function; therefore, the confidence ellipses for small changes in TAR and FAR appear as ellipses both on a linear and a semilog graph.)

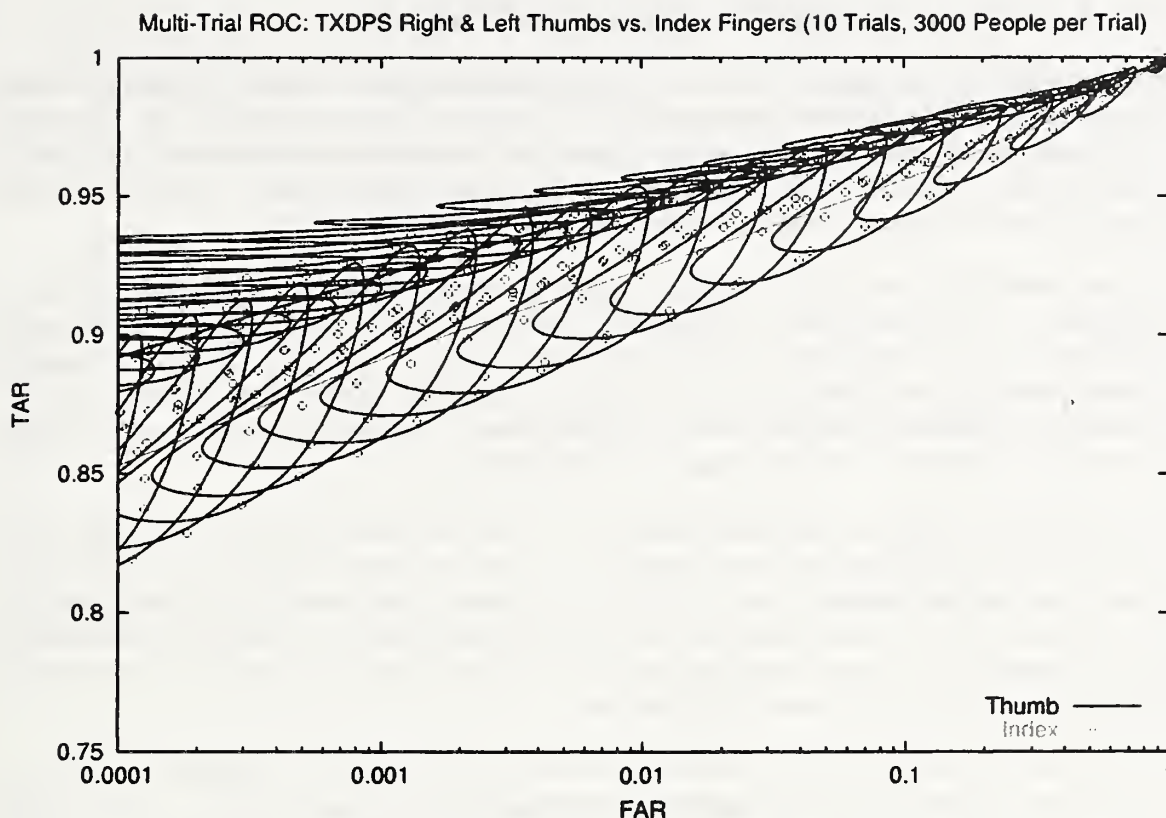


Figure 17. TXDPS Thumb vs. Index Finger Large Scale Verification –Multi-Trial ROC

TXDPS	
FAR @ 1% & TAR @ 98%	
Right & Left Thumb	
(1%, 95%)	(24%, 98%)
Right & Left Index	
(1%, 92%)	(40%, 98%)

Table 8. TXDPS Thumb vs. Index Results

A study was conducted whereby the two sets of ten 6K×6K similarity matrices were divided into upper-left and lower-right quadrants, and a Multi-Trial ROC analysis was separately performed on each set of submatrices. The results were similar to the analysis conducted on the DOS repository and plotted in Figure 15. The submatrix curves closely overlapped with the full matrix curve.

5.10 Large Scale Identification Study with DHS2

The experiments documented in this report to this point have been verification studies in which the application is simulated of determining if a person is who he claims to be. A single fingerprint is presented (a probe image) and determined to match a single enrolled fingerprint (a gallery image). This is a key part of a biometrically enabled border control system and perhaps represents the vast volume of processing that takes place in such a system.

It is also important to be able to accurately identify a person from an existing large repository of fingerprints. This is particularly important for enrollment where a person is not to be issued more than one travel document or card. In this case, a person's fingerprint(s) must be searched against a potentially very large repository in order to determine if this person is already enrolled, or perhaps to determine if this person has a criminal history. In practice, identification will likely involve searching not just one large repository, but rather several.

An identification study was designed using DHS2. Pairs of right and left index fingers were selected from the ~600K people in the repository. For the purposes of this study, the right index fingers were matched separately from the left index fingers. The second fingerprint impression from each pair of right index fingers was added to one large gallery of ~600K people, while the second fingerprint impression from each pair of left index fingers was added to a second large gallery corresponding to the same ~600K people.

The list of the ~600K people was randomly shuffled and was used to determine the order in which probe images were to be selected and matched against an entire gallery. Probe images were processed in blocks of 100 people. The results from ten blocks of random probes (a total of 1000 people) were matched and rank-1 statistics tabulated.

For identification, one of the most common methods of analyzing performance is the percentage of correct identification achieved at rank-1. In other words, when probes are matched to every fingerprint in the gallery, how often is the highest scoring match produced by that probe fingerprint and a gallery fingerprint belonging to the same person?

Figure 18 plots this statistic for ten random blocks of people from DHS2. For each block, a percentage is reported for right index finger results, and a second percentage is reported for left index finger results. Note the variation between blocks of people, and the variation at times between right and left index fingers within the same block.

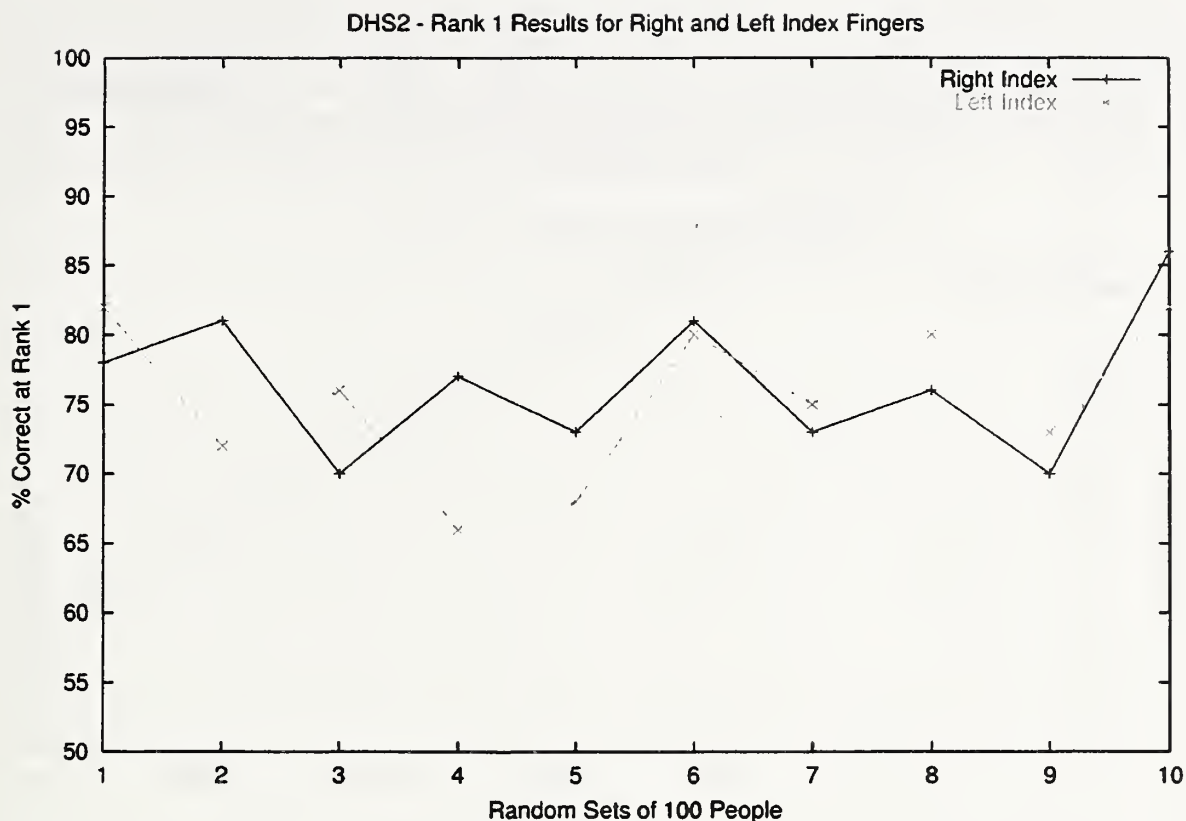


Figure 18. DHS2 Right & Left Index Finger Identification - % Correct by blocks of 100 people

There are 1000 random people represented in the figure above, and while these results are interesting, a more aggregate analysis of performance is desired. For example, if the size of the gallery changes, what effect might there be on system performance? A representation of rank-based performance in terms of gallery size is depicted in Figure 19.

In this figure, increasing gallery size is represented in log scale along the x-axis, while the percentage of correct identifications at a specific size of gallery is plotted on the y-axis. As can be seen, correct identification steadily decreases as the size of the gallery grows. The right-most point on each curve reports the identification rate achieved when the gallery contains all the people in the DHS2 repository. Also note that right index fingers consistently perform only slightly better than do left index fingers. With a nearly 600K gallery, an identification rate of 76% is achieved with right index fingers, while 75% is achieved with left index fingers.

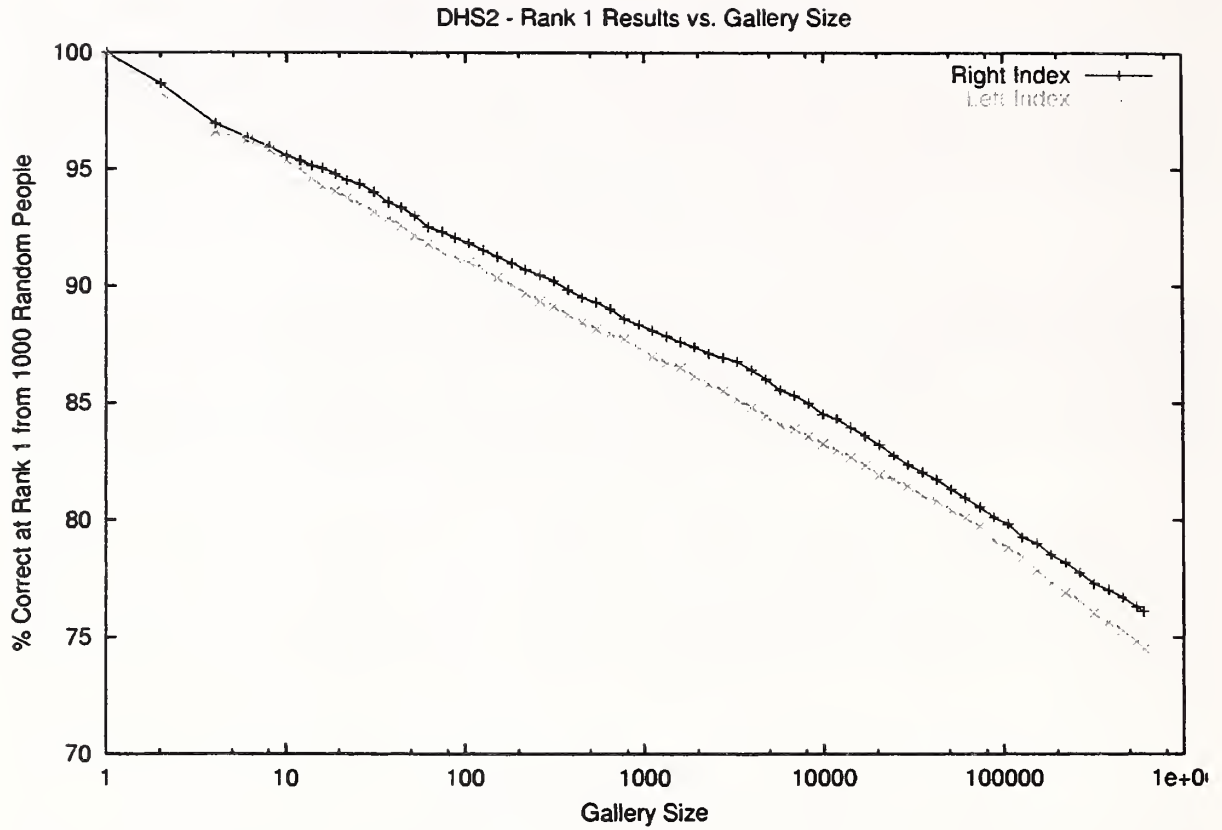


Figure 19. DHS2 Large Scale Identification – Comparison of right and left index fingers

5.11 Large Scale Identification Study with DOS

The same identification study in the previous section was conducted on the DOS repository. 600K pairs of right and left index fingers were selected from the repository, and right index fingers were matched and analyzed separately from left index fingers. The first impression from each pair of index fingers was added to its corresponding right or left index finger gallery set.

The list of 600K people was randomly shuffled and was used to determine the order in which probe images were to be selected and matched against an entire gallery. Probe images were processed in blocks of 100 people. The results from ten blocks of random probes (at total of 1000 people) were matched and rank-1 statistics tabulated.

Figure 20 plots the percentage of correct identifications achieved at rank-1 for each of the ten blocks of 100 probe images. For each block, two percentages are reported. The red curve is from right index fingers, while the green curve is from left index fingers. Note right index fingers consistently perform better than or equal to left index fingers.

The average identification rate across 1000 random DOS people is 82% for right index fingers and 72% for left index fingers. This contrasts with the DHS2 results, where average performance was 76% and 75% for 1000 random people and a comparably sized gallery set. Interestingly, the right index fingers from DOS perform better than those from DHS2, while left index fingers from DOS perform worse than those from DHS2.

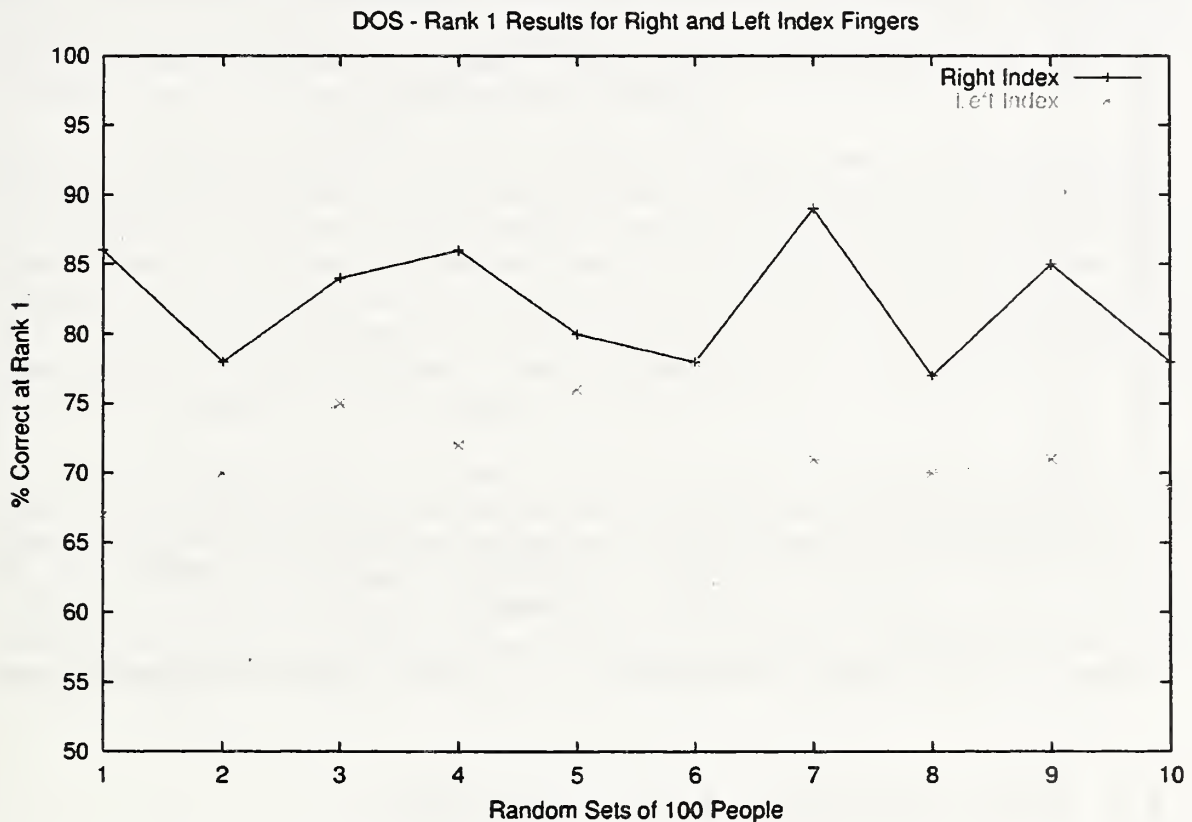


Figure 20. DOS Right & Left Index Finger Identification - % Correct by blocks of 100 people

An aggregate analysis depicting the effect increasing gallery size has on identification performance is plotted in Figure 21. The top, red curve is from right index fingers, and the bottom, green curve is from left index fingers. Notice that there is a significant separation between the right and left index finger results.

Comparing the identification results on DOS fingerprints in Figure 21 with results from DHS2 fingerprints in Figure 19, once again shows that DOS right index fingers perform better, but DOS left index fingers perform worse.

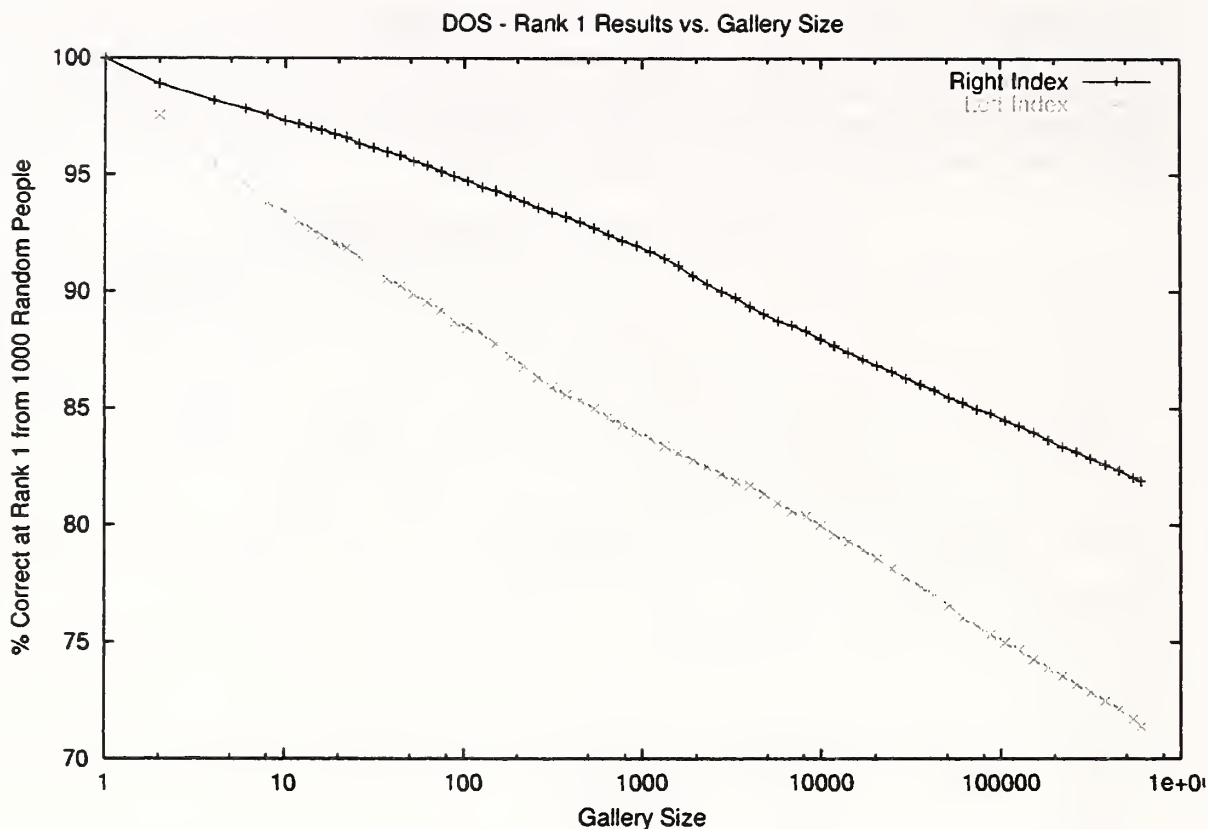


Figure 21. DOS Large Scale Identification – Comparison of right and left index fingers

5.12 Fusion of Results from Multiple Fingerprints

5.12.1 Score-Based Fusion Using SD29

A very interesting question to explore is, “What effect does combining matcher scores from multiple fingers have on performance?” A simple analysis was conducted to look at what effect there might be if matcher scores from thumbs were combined with those from index fingers. This is depicted by the scatter plot shown in Figure 22.

In this figure, plain matched to plain scores from SD29 thumbs are plotted along the x-axis, and plain matched to plain scores from corresponding (comparison from the same probe person and gallery person) index finger are plotted along the y-axis. The red ‘+’ points represent match scores, where the probe and gallery are from the same person. These points (~400) are labeled “Mate Scores” to avoid the ambiguity of the word “match.” The green ‘x’ points represent non-match scores, where the probe and gallery are from different persons. These (~90k) points are labeled “Non-Mate Scores” in the plot. As expected, the match scores are generally higher than the non-match scores. Note that the majority of green, non-match, points is clustered tightly at the origin.

First, examine the potential of utilizing only thumb scores. To accomplish this, one must decide a scalar threshold along the x-axis, at which point a vertical line is drawn along the graph and those points to the right of the threshold line are automatically assigned to the match class, while all points to the left of the threshold line are automatically assigned to the non-match class. The vertical dashed line labeled “Thumb Thresh” in the graph is a reasonable threshold on thumb scores, as nearly all the green cluster of points lies to the left of this threshold. In this case, notice the few green points to the right of this threshold. These green points represent potentially incorrect system identifications. Notice the large number of red points to the left of this threshold. These red points represent missed identifications.

Second, examine the potential of utilizing only index finger scores. For index fingers, a horizontal threshold line must be determined. The horizontal dashed line labeled “Index Thresh” in the graph is a reasonable threshold on index finger scores, as nearly all the green cluster of points lies below this threshold. Once again, there are a few green points above this threshold, representing potentially incorrect system identifications. Notice that there are more red points below the index finger threshold than there are red points to the left of the thumb threshold. Index finger scores will cause a greater number of identifications to be missed when the uniform threshold value in this illustration is applied to both axes.

Now look at the diagonal dashed line labeled “Combined Thresh” in the graph. This represents a simple linear threshold based on combining both the thumb score and the index finger score. Notice the tight fit of this threshold along the edge of the green cluster of points. The majority of the red points that were missed by the previous one-finger thresholds are now accurately separated (up and to the right) from the green cluster. There are some red points that still overlap with the green cluster, and these are cases that may be separated by combining scores from additional fingers.

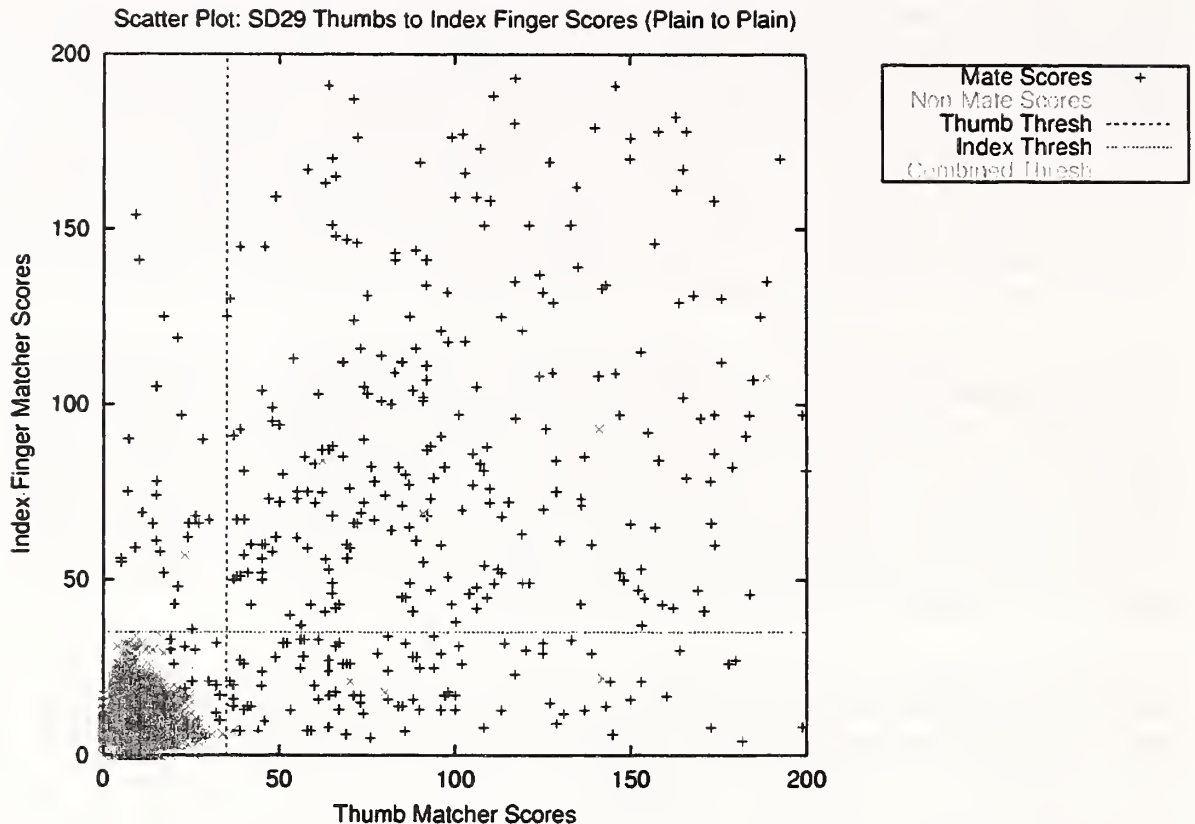


Figure 22. SD29 Combined Thumb & Index Finger Scores

Given the observations from the scatter plot above, a 2-finger ROC analysis was conducted. A linear threshold was used with slope = -1, in which case the combined matcher score is simply calculated as $S_c = S_t + S_i$; where S_c is the combined score, S_t is the matcher score for the thumb and S_i is the matcher score for the index finger. This threshold was incrementally applied across the range of 2-finger points of (thumb, index) matcher scores. At each increment of linear threshold, true accepts were accumulated from those match distribution points (red '+'s) remaining above the threshold line while false accepts were accumulated from those non-match distribution points (green 'x's) remaining above the threshold line. The illustration in Figure 22 shows the linear threshold (labeled "Combined Thresh") at an x and y intercept of 50.

By sweeping this simple linear threshold across the range of points, the top ROC curve in Figure 23 was computed. The middle green curve in the figure is the plain matched to plain results for SD29 thumbs from Figure 5. The lower blue curve in the figure is the plain matched to plain results for SD29 index fingers from Figure 6. The fused red curve shows a substantial improvement in performance.

As can be seen in Figure 23, at a false accept rate of 1%, the true accept rate from thumbs improved by 80% (95% to 99%) by adding index finger scores to the decision. The combined results are quite remarkable, especially in light of the fact that the true accept rate of the index fingers was only 90% at the same false accept rate of 1%.

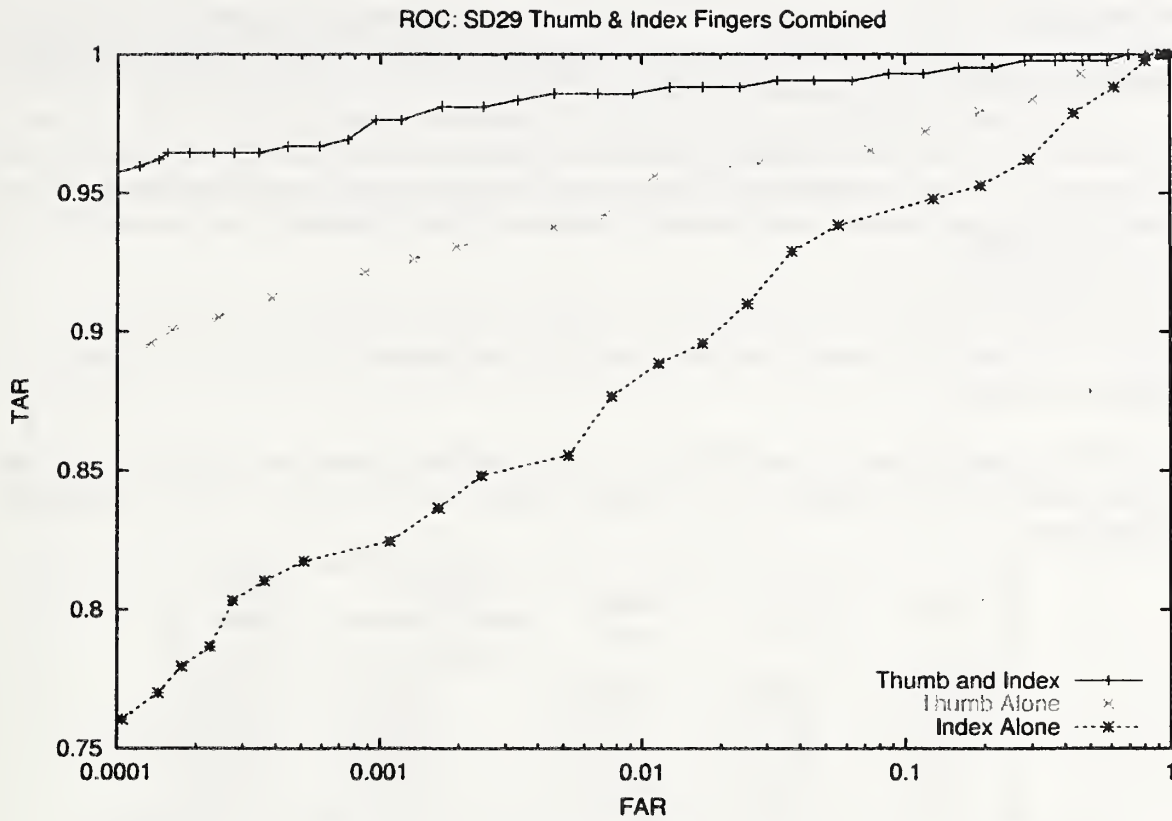


Figure 23. SD29 Combined Thumb & Index Finger Verification

SD29 Thumb w/ Index	
FAR @ 1% & TAR @ 98%	
Thumb & Index Combined	
(1%, 99%)	(0.2%, 98%)
Thumb Alone	
(1%, 95%)	(19%, 98%)
Index Alone	
(1%, 90%)	(43%, 98%)

Table 9. SD29 Score-Based Thumb and Index Finger Fusion Results

The theoretical limits of score-based fusion can be easily calculated. If the scores to be fused are statistically not independent, then the fused performance may be better than either of the original

scores, but will not achieve the ideal case. If the scores to be fused are statistically independent, then for each subject, the probability that neither finger will match is equal to the product of the individual probabilities of not matching (the ideal case).^{††}

In practice, this means that the ideal performance of fused scores can be predicted, assuming that the scores to be fused are statistically *independent*. Comparing the actual fused results to the predicted ideal performance can provide a measure of the statistical independence of the constituent scores.

For this purpose, TAR (True Accept Rate) should be stated as FRR (False Reject Rate) = 1 – TAR. If the index and thumb scores are statistically independent, for a given FAR, $FRR_{fused} = FRR_{index} * FRR_{thumb}$,

The following table shows that the ROC for the combined index fingers and thumbs is very close to the prediction, and therefore the constituent scores are very close to being statistically independent.

FAR	Index TAR	Thumb TAR	Actual Index*Thumb TAR	Ideal Fused TAR
0.0001	76.0%	89.5%	95.7%	97.5%
0.001	82.5%	92.5%	97.6%	98.7%
0.01	88.8%	95.6%	98.6%	99.5%
0.1	94.7%	97.2%	99.3%	99.9%

Table 10. Actual and Ideal Score-Based Fusion for Index Fingers and Thumbs

While the effect of combining thumb and index fingers together is dramatic, there remains a more practical question regarding the effect if right and left right index fingers are combined. Current border control and visa systems are capturing index fingers, not thumbs. A study was conducted to examine the effect of combining right and left index fingers.

Due to the small size of SD29 (216 people, of whom only 207 had complete pairs of index fingers), the results reported above on index fingers included match scores among both right and left index fingers. To study the effect of combining the match scores from a person's right and left index fingers, the right matched to right index finger scores were separated from left matched to left index finger scores. This resulted in two similarity matrices each of size 207×207, one for right index fingers and one for the left.

Figure 24 contains a scatter plot with plain matched to plain scores from SD29 right index fingers plotted along the x-axis, and plain matched to plain scores from corresponding (comparison from the same probe person and gallery person) left index finger are plotted along the y-axis. The red '+' points represent match scores, where the probe and gallery are from the same person. These points (207) are labeled "Mate Scores." The green 'x' points represent non-

^{††} Mitretek Systems, *Image Quality Study*, Dec. 2000, pg. 6-20.

match scores, where the probe and gallery are from different persons. These (~43k) points are labeled “Non-Mate Scores” in the plot. These results are strikingly similar to those for thumbs combined with index fingers in Figure 22. The match scores are generally higher than the non-match scores with the majority of green, non-match, points clustered tightly at the origin.

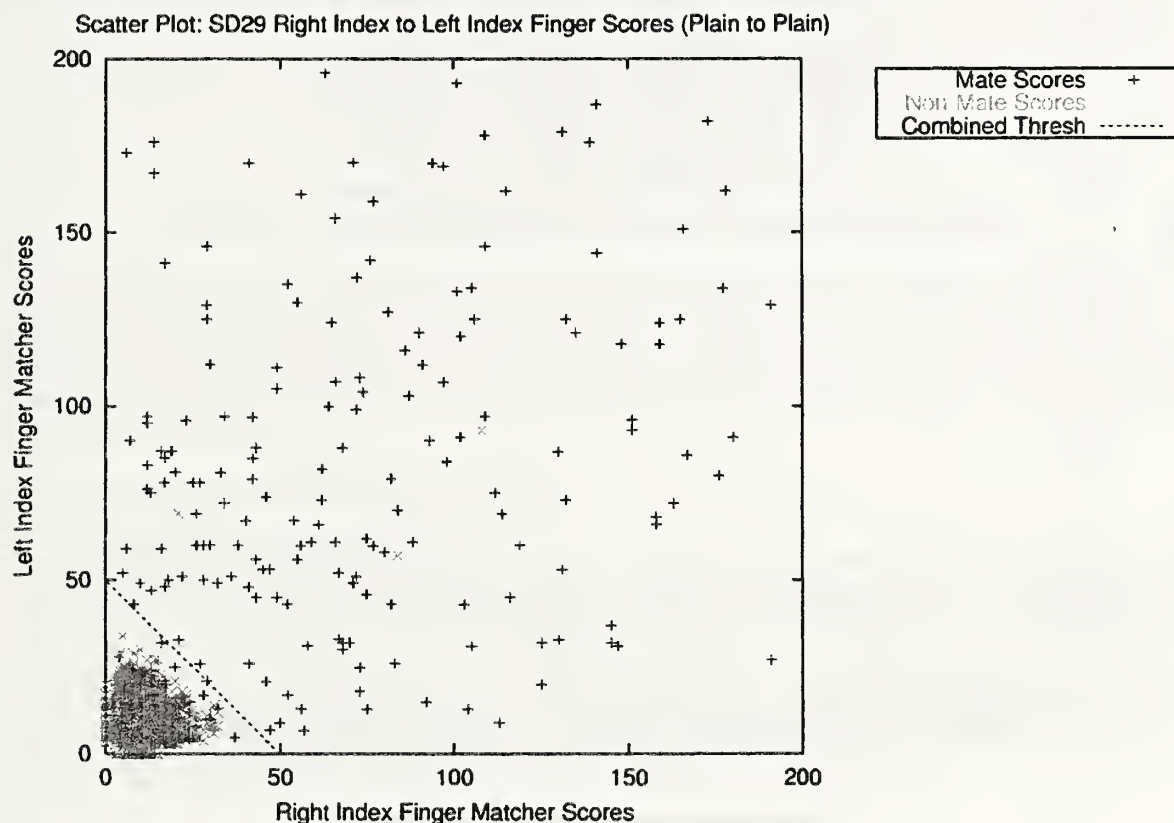


Figure 24. SD29 Combined Right & Left Index Finger Scores

The top ROC curve in Figure 25 was computed by applying a linear threshold with slope = -1 (in which case the combined matcher score is simply calculated as $S_c = S_r + S_l$; where S_c is the combined score, S_r is the matcher score for the right finger and S_l is the matcher score for the left finger) across the range of scores in the scatter plot above. The lower green curve belongs to the results of individually matching each index finger (right and left) to all other index fingers (right and left). This is the same curve labeled “Index Alone” in Figure 23. It was the matcher scores that comprise this composite curve that were separated into right index finger and left index finger similarity matrices and then combined to produce the upper curve labeled “Right Index and Left Index” in Figure 25.

At a FAR of 1%, the combined index fingers performed (98%) 1% lower than the combined thumb and index finger (99%).

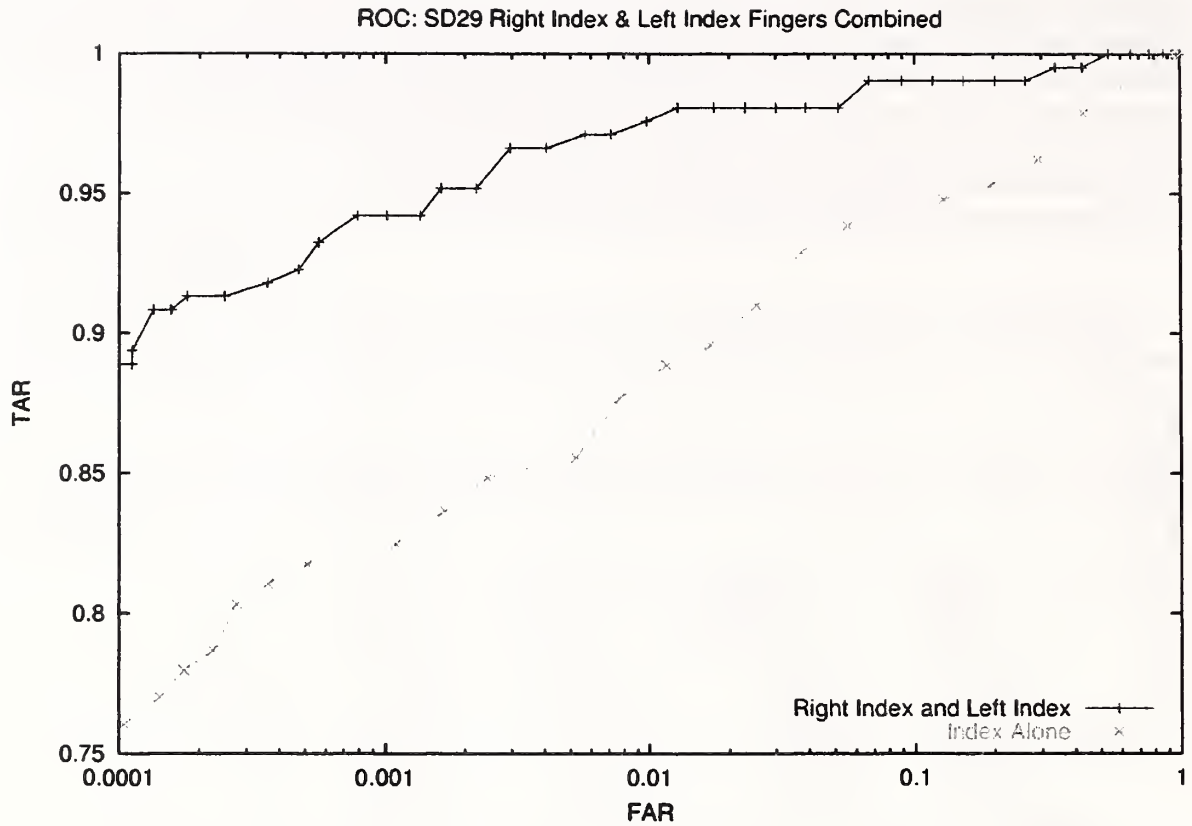


Figure 25. SD29 Combined Right & Left Index Finger Verification

SD29 Index w/ Index	
FAR @ 1% & TAR @ 98%	
Right & Left Index Combined	
(1%, 98%)	(1%, 98%)
Index Alone	
(1%, 90%)	(64%, 98%)

Table 11. SD29 Score-Based Right and Left Index Finger Fusion Results

Using the same method of predicting ideal fusion performance as in the index-thumb fusion case, the results again show that the constituent scores are very close to being statistically independent.

FAR	Separate Index TAR	Fused Index TAR	Ideal Fused TAR
0.0001	76.0%	88.9%	94.2%
0.001	82.5%	94.2%	96.9%
0.01	88.8%	97.6%	98.7%
0.1	94.7%	99.0%	99.7%

Table 12. Actual and Ideal Score-Based Fusion for Index Fingers

At very high FAR levels (such as 1%), near perfect verification results can be achieved by score-based fusion of two fingers. For a large scale system that uses score-based thresholds (such as an AFIS), the fusion of multiple fingers is fundamental to performance accuracy. This is why at extremely low FAR levels (such as below 10^{-8}), more than two fingers must be used to achieve reasonable TAR levels.

5.12.2 Score-Based Fusion Using DHS10-C

The same method of score-based fusion used in Section 5.12.1 was conducted on the 1021 people in DHS10-C. The top red ROC curve in Figure 26 shows the results of combining the matcher scores of a person's right thumb and index finger. The middle green curve plots the results of using the right thumb alone, while the bottom blue curve plots the results of using right index fingers alone. As was the case with SD29, significant improvement with DHS10-C is achieved when thumb and index finger are combined; however, at a FAR of 1%, the fused performance with DHS10-C is 4% lower (95% vs. 99%) than with SD29.

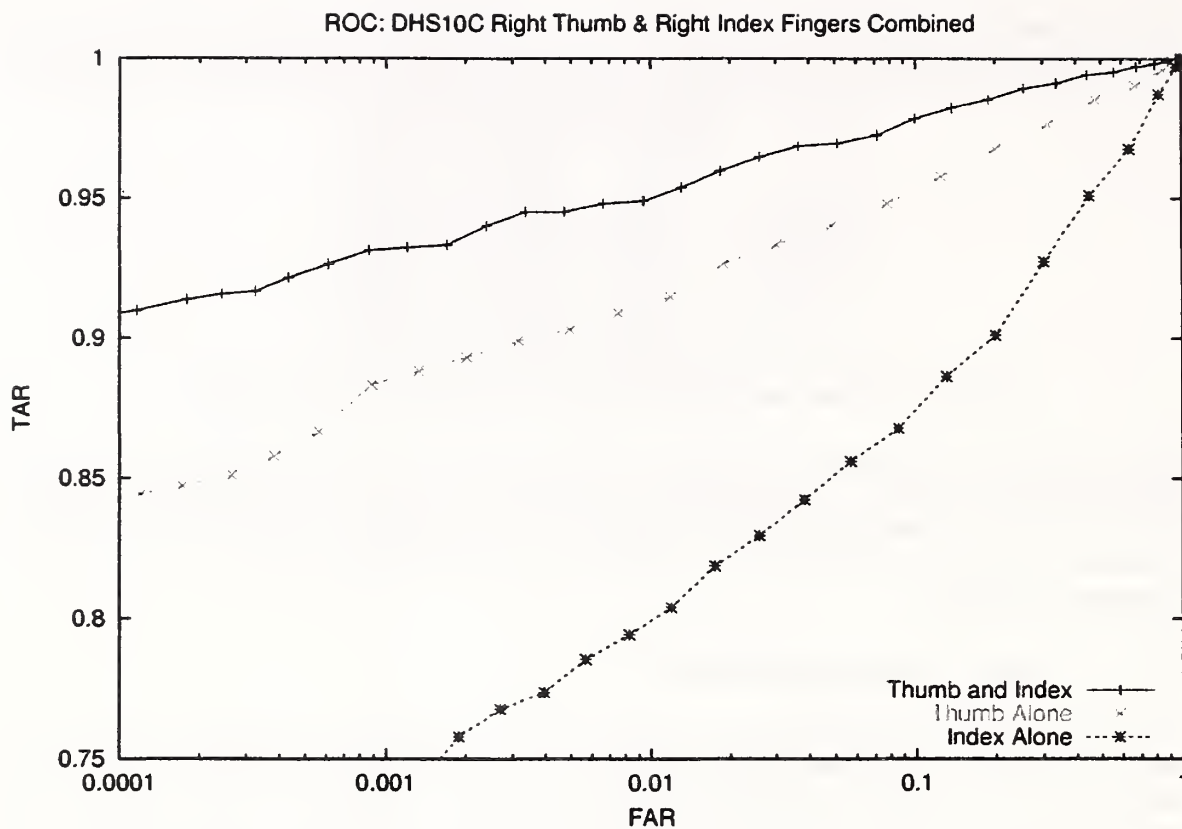


Figure 26. DHS10-C Combined Right Thumb & Right Index Finger Verification

DHS10-C Right Thumb w/ Index	
FAR @ 1% & TAR @ 98%	
Right Thumb & Index Combined	
(1%, 95%)	(13%, 98%)
Right Thumb Alone	
(1%, 91%)	(39%, 98%)
Right Index Alone	
(1%, 80%)	(80%, 98%)

Table 13. DHS10-C Score-Based Right Thumb and Right Index Finger Fusion Results

Figure 27 shows the fusion results with DHS10-C when combining matcher scores of right and left index fingers. Again, the fused performance is dramatically higher than either of the index finger results alone. At a FAR of 1%, the combined index fingers performed (92%) 3% lower than the combined right thumb and right index finger (95%).

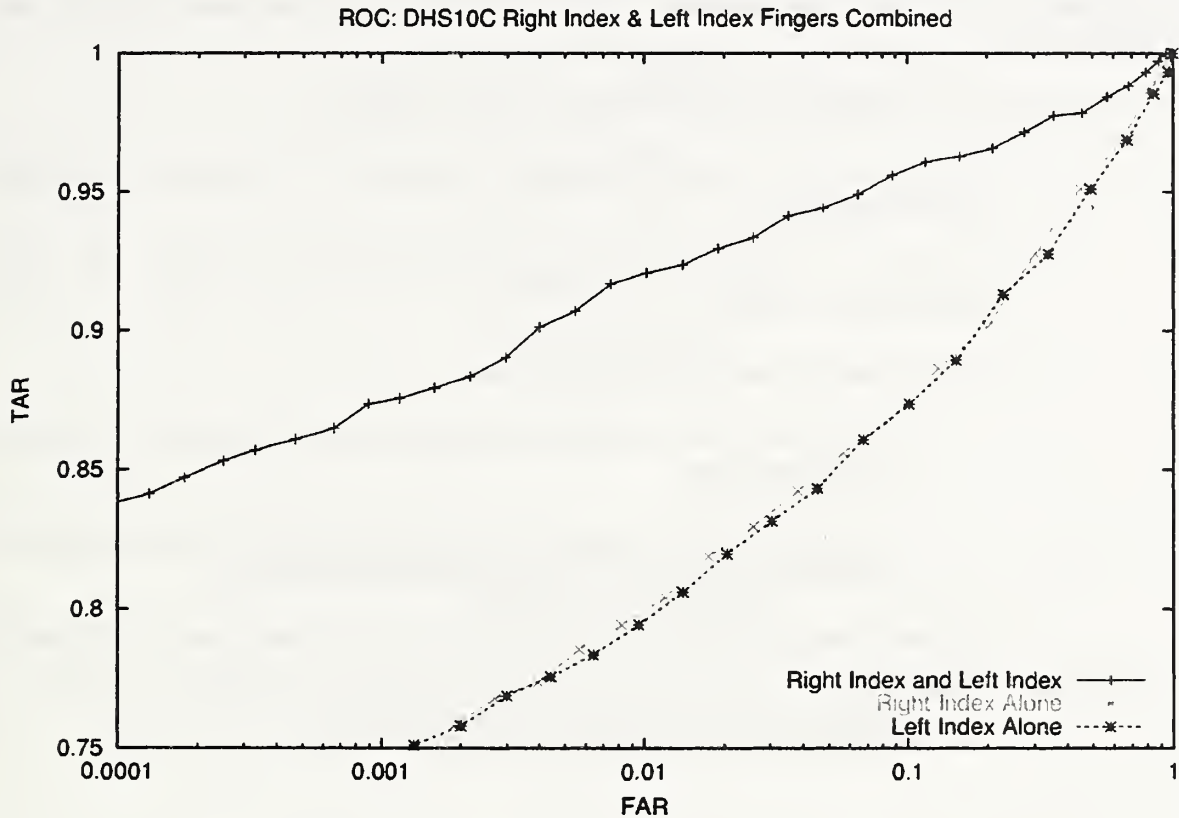


Figure 27. DHS10-C Combined Right & Left Index Finger Verification

DHS10-C Index w/ Index	
FAR @ 1% & TAR @ 98%	
Right & Left Index Combined	
(1%, 92%)	(49%, 98%)
Right Index Alone	
(1%, 80%)	(80%, 98%)
Left Index Alone	
(1%, 80%)	(80%, 98%)

Table 14. DHS10-C Score-Based Right and Left Index Finger Fusion Results

5.12.3 Rank and Score-Based Fusion Using DHS2

As discussed in Section 5.12.1, a significant improvement in score-based performance may be obtained by combining matcher scores from two different fingerprints of the same person. A simple algorithm was designed and tested to determine the effect adding a second finger has on rank-based identification performance. The following algorithm was proposed and implemented:

- Given a probe person to be identified from a gallery of people ...
 - Match the probe person's right index fingerprint to all fingerprints in the right index finger gallery.
 - Compile a list of gallery people associated with the top-100 highest right index finger matcher scores.
 - For each gallery person in the top-100 list
 - Match probe person's left index fingerprint with the gallery person's left index fingerprint.
 - Add the right index finger score and the left index finger score from matching the probe person to the current gallery person
 - All other combined scores belonging to gallery persons not in the top-100 list are set to zero.
 - The highest combined score is deemed the rank-1 selection for the probe

The decision to only combine the left index finger scores with the top-100 right index fingers scores is computationally strategic. The bulk of the time is spent matching the probe person's right index fingerprint to large number of gallery fingerprints. Combining the second finger only costs an additional sort and 100 additional matches. Thus the second finger is added with little computational overhead.

This algorithm was tested using the same 1000 random people used in the previous DHS2 identification study. Figure 28 illustrates the effect of combining right and left index finger scores. Identification rates are reported for ten random blocks of 100 people. The lower curve is the result of using right index finger scores alone. The upper curve is the result of adding the left index finger scores according to the algorithm.

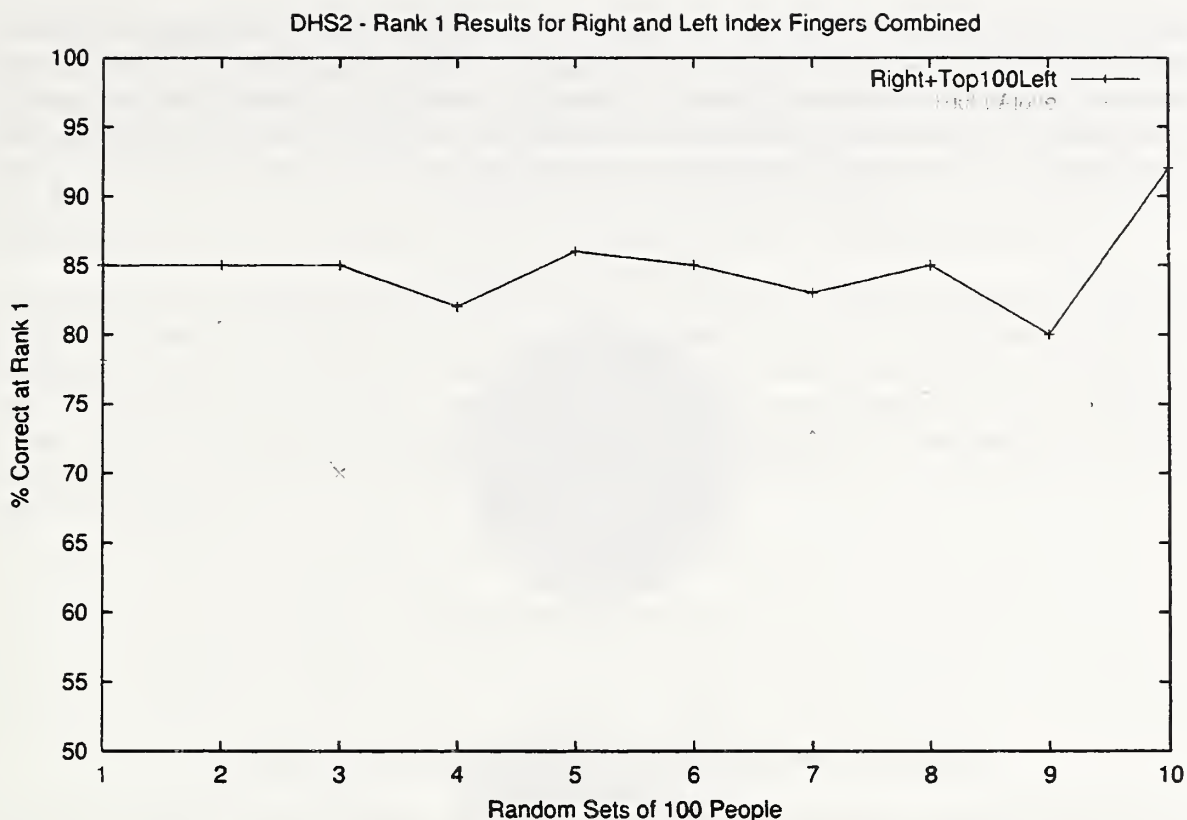


Figure 28. DHS2 Combining Right and Left Index Finger Identification – Comparing combined results with right index finger alone

The decision to use a threshold of the top-100 was made empirically by looking at histograms of right index finger scores. If the matcher score between the probe person's right index fingerprint and its mate in the gallery is not sufficiently high to make it in the list of top 100 scores, then it is impossible for the algorithm to improve identification. An upper bound for the algorithm is determined by accumulating correct identifications within the top-100 right index finger scores alone. (This is referred to as a cumulative match score at rank-100.)

The aggregate identification rate across 1000 random people searched against ~600K DHS2 gallery increased from 76% to 85% when the left index finger score was combined with the right index finger score in the identification decision, whereas the cumulative match score at rank-100 for the right index finger scores alone was 86%. Based on the above results, the algorithm improved identification by 9% and nearly met its potential in that the difference between the realized identification rate and the upper bound was only 1%.

An in-depth analysis was conducted to study how the algorithm performed at several key stages. This is represented by a series of pie charts that follow. Figure 29 illustrates two sets of scores from matching the probe person's right index fingerprint to its mate in the gallery. From 1000

probe persons, those scores in blue on the left (76%) labeled “Rank-1” are matches at rank-1, resulting in the highest score across the entire gallery. These represent correct identifications when using right index finger scores alone. Those scores in burgundy on the right (24%) labeled “Not Rank-1” are matches to gallery mates not resulting in the top score. These represent missed identifications when using right index finger scores alone.

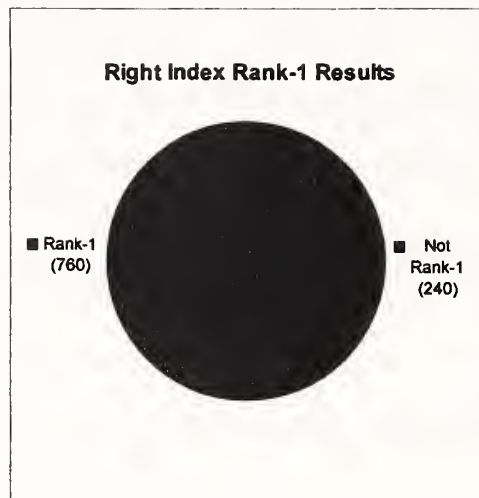


Figure 29. Identification Using Right Index Finger Alone

As mentioned before, if the right index finger score is not within the top-100 of all gallery scores, then the algorithm has no chance of correcting the missed identification. This is represented in Figure 30, where the burgundy set in the previous chart is now subdivided into two categories. Right index finger scores ranking within the top-100 scores across the entire gallery (10.7%) remain in the top burgundy set labeled “Top-100”, while those not making it in the top-100 (13.3%) are in the bottom ivory set labeled “Not In Top-100” and represent identifications that cannot be aided by the algorithm.

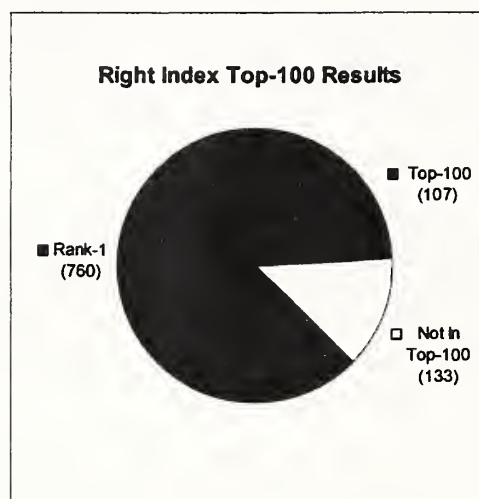


Figure 30. Right Index Fingers Scoring in Top-100

Those cases represented in the burgundy set in the previous figure have the potential of being moved to rank-1 when the left index finger score is added with their right index finger score. In Figure 31, the burgundy set labeled “Move To Rank-1” represents those cases where identification, missed based on the right index finger score alone, is now correctly made at rank-1 when left and right index finger matcher scores are combined. The orange set labeled “Stay Not Rank-1” represents those cases that remain missed identifications. Of the 107 cases possible, 89 move to rank-1 while only 18 remain missed identifications for a yield of 83%.

The algorithm also adds left index finger scores to the 760 cases (blue set in the previous chart) that achieve correct identification with the right index fingers alone. In a situation where the resulting matcher score of the probe person’s left finger to its mate in the gallery is particularly bad, it is possible for the combined score to sufficiently decrease relative to other combined scores so that the identification moves from correct to being missed. Fortunately, this only happened to 2 of the 760 cases and is represented by the light blue wedge labeled “Move To Not Rank-1” in Figure 31.

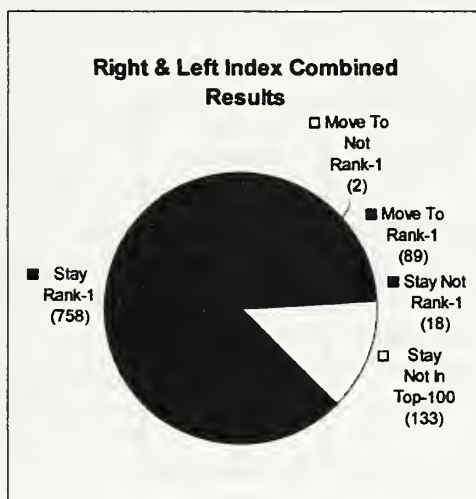


Figure 31. Identification Combining Left with Right Index Finger Scores

In summary, a simple algorithm was designed and tested to study the effect of adding a second finger to the identification decision. By adding left index finger scores to just the top-100 right index finger scores, no significant time in computation was added and missed identifications were reduced by more than 37% (9 / 24). Note that an even greater yield could be achieved if the threshold in the algorithm were to be increased from top-100. This would result in a smaller ivory set labeled “Stay Not In Top-100” in Figure 31, but at a greater computational cost.

5.13 Person Variation Study with DHS2

The study reported in this section explores whether some people’s fingerprints are intrinsically more difficult to match than others. Obviously, some fingerprints are more difficult to match than others, but this section explores the extent to which quality may be inherent in the person’s finger itself, rather than in a specific instance of a fingerprint image.

The DHS2 repository contains a minimum of two pairs of fingerprints per person, but in a number of cases people have many more impressions of their index fingers. Having this data permits the analysis of the variation of matcher performance between fingerprints from the same person (referred to as *intra-person* variation) as well as variation of matcher performance between fingerprints from different people (referred to as *inter-person* variation). A verification experiment was designed in order to look at these types of person-level variations and to look for ways to identify people whose fingerprints are more difficult to match than others, and reasons then as to why could be explored.

One hundred people were selected from DHS2, each having at least 30 right index finger impressions in the repository. Selecting the first 30 impressions from each person resulted in a list of 3000 fingerprints, and using these same 3000 fingerprints as a probe list and a gallery list, a 3K×3K similarity matrix of matcher scores was computed as illustrated in Figure 32.

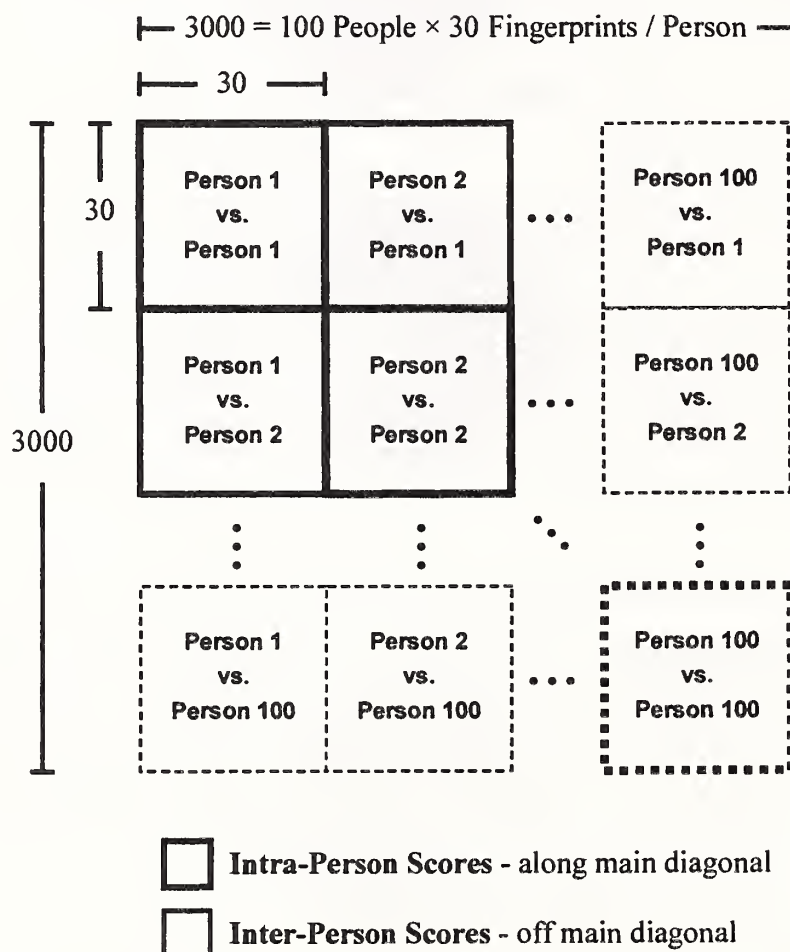


Figure 32. Composition of Similarity Matrix for Person Variation Study

Looking at the illustration, the similarity matrix was organized by person. Probe fingerprints were represented along the horizontal dimension while gallery fingerprints were represented along the vertical dimension. Along the horizontal dimension, the 30 right index fingerprints belonging to the first person were followed by the 30 right index fingerprints of the second person, and so forth. The same order was used down the vertical dimension in such a way so as to subdivide the similarity matrix into a grid of person-compared-to-person blocks of fingerprint scores.

The 30×30 blocks running down the main diagonal, colored with a blue border in the illustration, represent intra-person matcher scores, where the 30 fingerprints belonging to the same person were fully matched among themselves. The black bordered 30×30 blocks off the main diagonal represent inter-person matcher scores, where the 30 fingerprints from one person were fully matched to the 30 fingerprints of a different person. In this way, trends as to the difficulty of a particular person as well as a particular fingerprint could be studied.

Once the similarity matrix was computed, analysis began by examining the intra-person results within the blocks down the main diagonal. An example of the fingerprints belonging to one particular person (Person 9) is shown in Figure 33. These would be characteristic of good quality fingerprints in DHS2. The images have reasonably good contrast, with good definition between friction skin ridges and valleys, which supports reliable minutiae extraction and matching.

The block of matcher scores for this good quality example are listed in Figure 34. Using the Bozorth98 matcher, scores will range from 0 to 499, and sometimes higher. Scores over 40 typically represent a match, meaning the two fingerprints were from the same finger from the same person. Notice that the scores down the main diagonal of the block are significantly higher. These are the result of matching an impression with itself, and should be treated separately from those scores off the main diagonal. The mean of the off-diagonal scores for this good quality example is 98.



Figure 33. Good Quality Fingerprints from DHS2 Person 9

499.0	131.0	102.0	48.0	99.0	64.0	114.0	106.0	77.0	76.0	65.0	39.0	67.0	68.0	54.0	83.0	93.0	87.0	21.0	37.0	68.0	72.0	93.0	98.0	137.0	103.0	136.0	83.0	68.0	86.0
130.0	476.0	133.0	85.0	190.0	73.0	122.0	110.0	114.0	92.0	81.0	52.0	93.0	136.0	74.0	130.0	93.0	115.0	33.0	75.0	87.0	86.0	105.0	87.0	152.0	158.0	101.0	94.0	127.0	122.0
102.0	135.0	499.0	117.0	210.0	137.0	104.0	81.0	121.0	108.0	134.0	64.0	112.0	161.0	83.0	80.0	83.0	102.0	44.0	123.0	186.0	117.0	135.0	144.0	152.0	179.0	125.0	129.0	124.0	129.0
48.0	86.0	117.0	499.0	115.0	96.0	104.0	77.0	67.0	46.0	95.0	35.0	56.0	59.0	43.0	84.0	94.0	98.0	71.0	95.0	71.0	56.0	76.0	83.0	105.0	85.0	67.0	71.0	101.0	68.0
107.0	196.0	206.0	129.0	499.0	84.0	179.0	152.0	182.0	160.0	123.0	96.0	195.0	246.0	123.0	126.0	150.0	215.0	56.0	87.0	114.0	176.0	165.0	166.0	129.0	172.0	154.0	159.0	179.0	179.0
63.0	69.0	137.0	105.0	88.0	499.0	119.0	65.0	85.0	52.0	66.0	37.0	34.0	53.0	34.0	78.0	101.0	61.0	39.0	76.0	123.0	50.0	64.0	59.0	114.0	90.0	55.0	81.0	74.0	62.0
114.0	131.0	151.0	102.0	180.0	119.0	499.0	149.0	159.0	116.0	98.0	75.0	98.0	126.0	66.0	110.0	116.0	128.0	39.0	68.0	135.0	135.0	154.0	169.0	196.0	185.0	142.0	167.0	107.0	128.0
106.0	109.0	81.0	69.0	134.0	65.0	137.0	473.0	111.0	98.0	69.0	63.0	118.0	97.0	94.0	88.0	83.0	144.0	49.0	51.0	75.0	121.0	123.0	123.0	88.0	115.0	119.0	81.0	103.0	106.0
78.0	116.0	121.0	69.0	182.0	85.0	159.0	112.0	499.0	138.0	101.0	77.0	116.0	134.0	109.0	107.0	129.0	115.0	48.0	78.0	92.0	105.0	143.0	132.0	101.0	108.0	137.0	144.0	100.0	130.0
76.0	93.0	109.0	47.0	160.0	52.0	115.0	98.0	136.0	499.0	72.0	70.0	89.0	103.0	97.0	72.0	85.0	82.0	52.0	51.0	92.0	75.0	105.0	92.0	98.0	97.0	113.0	121.0	61.0	90.0
65.0	82.0	134.0	95.0	122.0	68.0	99.0	67.0	102.0	68.0	499.0	50.0	84.0	88.0	76.0	88.0	87.0	113.0	49.0	67.0	73.0	81.0	112.0	111.0	101.0	108.0	126.0	95.0	104.0	82.0
39.0	55.0	64.0	38.0	98.0	36.0	75.0	65.0	77.0	73.0	51.0	499.0	76.0	92.0	84.0	49.0	58.0	61.0	15.0	46.0	73.0	70.0	95.0	84.0	39.0	92.0	62.0	76.0	47.0	84.0
67.0	93.0	111.0	55.0	195.0	34.0	98.0	118.0	115.0	89.0	82.0	75.0	353.0	172.0	83.0	77.0	72.0	123.0	32.0	60.0	84.0	140.0	125.0	112.0	81.0	109.0	91.0	79.0	105.0	146.0
73.0	143.0	160.0	59.0	232.0	53.0	127.0	97.0	132.0	103.0	80.0	98.0	171.0	480.0	94.0	77.0	81.0	129.0	31.0	66.0	124.0	162.0	134.0	142.0	105.0	149.0	88.0	110.0	116.0	138.0
54.0	74.0	83.0	43.0	124.0	34.0	65.0	94.0	110.0	97.0	76.0	84.0	86.0	91.0	453.0	65.0	58.0	89.0	22.0	58.0	68.0	87.0	102.0	97.0	68.0	75.0	92.0	87.0	66.0	83.0
88.0	129.0	80.0	85.0	127.0	77.0	110.0	88.0	107.0	71.0	89.0	51.0	77.0	79.0	65.0	499.0	109.0	98.0	48.0	52.0	63.0	65.0	93.0	92.0	82.0	99.0	85.0	74.0	101.0	79.0
93.0	93.0	82.0	92.0	150.0	108.0	117.0	85.0	130.0	82.0	72.0	58.0	72.0	81.0	58.0	124.0	499.0	142.0	46.0	74.0	74.0	92.0	98.0	87.0	128.0	92.0	87.0	80.0	128.0	121.0
90.0	113.0	96.0	97.0	209.0	60.0	124.0	155.0	107.0	86.0	112.0	61.0	126.0	117.0	87.0	101.0	148.0	413.0	49.0	66.0	77.0	130.0	133.0	87.0	119.0	132.0	105.0	90.0	144.0	138.0
21.0	33.0	44.0	71.0	56.0	39.0	39.0	49.0	48.0	52.0	49.0	17.0	33.0	31.0	22.0	48.0	46.0	49.0	519.0	48.0	29.0	30.0	51.0	37.0	27.0	38.0	28.0	37.0	53.0	40.0
39.0	75.0	123.0	95.0	87.0	75.0	69.0	51.0	79.0	54.0	70.0	45.0	60.0	66.0	58.0	52.0	75.0	66.0	48.0	499.0	73.0	57.0	64.0	50.0	88.0	85.0	77.0	69.0	78.0	55.0
75.0	87.0	185.0	73.0	117.0	123.0	134.0	77.0	89.0	92.0	73.0	77.0	86.0	125.0	68.0	60.0	76.0	80.0	29.0	71.0	499.0	137.0	74.0	106.0	125.0	144.0	88.0	97.0	66.0	122.0
65.0	86.0	116.0	55.0	173.0	50.0	135.0	123.0	109.0	75.0	81.0	76.0	141.0	162.0	88.0	65.0	101.0	131.0	30.0	57.0	137.0	410.0	141.0	103.0	124.0	106.0	81.0	119.0	92.0	154.0
93.0	106.0	135.0	72.0	167.0	64.0	154.0	113.0	140.0	105.0	117.0	94.0	126.0	134.0	101.0	93.0	98.0	134.0	51.0	63.0	79.0	141.0	481.0	179.0	118.0	132.0	156.0	109.0	117.0	134.0
98.0	87.0	144.0	80.0	166.0	59.0	158.0	123.0	137.0	100.0	111.0	82.0	112.0	142.0	98.0	88.0	82.0	83.0	36.0	52.0	105.0	103.0	178.0	499.0	105.0	144.0	212.0	106.0	73.0	89.0
153.0	152.0	141.0	101.0	137.0	114.0	191.0	88.0	108.0	96.0	101.0	41.0	81.0	105.0	72.0	93.0	130.0	120.0	27.0	84.0	125.0	127.0	114.0	103.0	474.0	191.0	117.0	122.0	99.0	115.0
105.0	162.0	180.0	83.0	172.0	90.0	185.0	113.0	117.0	98.0	109.0	94.0	109.0	149.0	78.0	101.0	85.0	130.0	38.0	87.0	134.0	106.0	132.0	145.0	182.0	499.0	121.0	143.0	111.0	149.0
136.0	102.0	133.0	71.0	160.0	52.0	144.0	121.0	135.0	113.0	125.0	62.0	91.0	86.0	96.0	80.0	87.0	106.0	28.0	72.0	89.0	81.0	156.0	213.0	116.0	122.0	487.0	116.0	83.0	98.0
83.0	94.0	126.0	77.0	160.0	79.0	170.0	76.0	145.0	117.0	97.0	79.0	80.0	112.0	88.0	78.0	80.0	90.0	37.0	70.0	89.0	118.0	109.0	105.0	130.0	144.0	119.0	471.0	79.0	139.0
71.0	127.0	123.0	100.0	179.0	75.0	105.0	103.0	100.0	61.0	104.0	49.0	106.0	112.0	66.0	102.0	127.0	143.0	53.0	77.0	67.0	92.0	123.0	72.0	90.0	112.0	82.0	79.0	437.0	129.0
87.0	120.0	128.0	68.0	167.0	62.0	128.0	106.0	130.0	90.0	81.0	84.0	147.0	146.0	82.0	79.0	121.0	137.0	40.0	55.0	120.0	155.0	134.0	90.0	115.0	146.0	97.0	139.0	130.0	499.0

Figure 34. Intra-Person Similarity Submatrix for Good Quality Case 1



Figure 35. Poor Quality Fingerprints from DHS2 Person 7

An example of characteristically poor quality fingerprints belonging to one particular person (Person 7) is shown in Figure 35. These images suffer from several factors including areas of low contrast, areas of smudging, and only partial areas of the fingerprint captured.

Figure 36 compares the distribution of matcher scores between the good quality example and the poor quality example. The scores for the poor quality case are significantly lower with a mean of only 18. Also notice that the shapes of the two distributions are quite different and easily distinguishable.

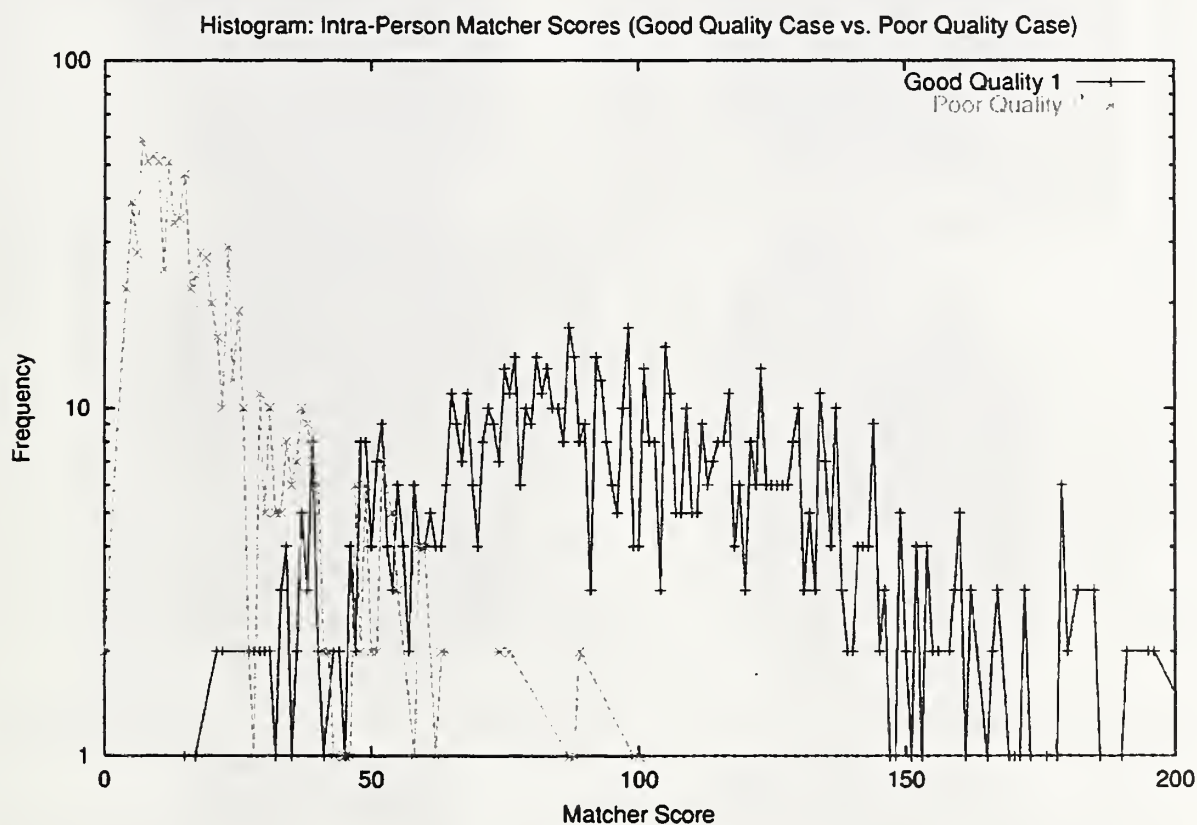


Figure 36. Comparison of Intra-Person Matcher Scores Between a Good and Poor Quality Case

The fingerprints from a second good quality example (Person 24) are shown in Figure 37. There appears to be a bit less contrast and more smudging in some of these fingerprints than in those of the first good quality example.



Figure 37. Good Quality Fingerprints from DHS2 Person 24

Figure 38 compares the distribution of matcher scores between the two good quality examples. As can be seen, the distributions are quite similar. This suggests that minutiae extraction and matching were not adversely affected by the image anomalies seen in Figure 37.

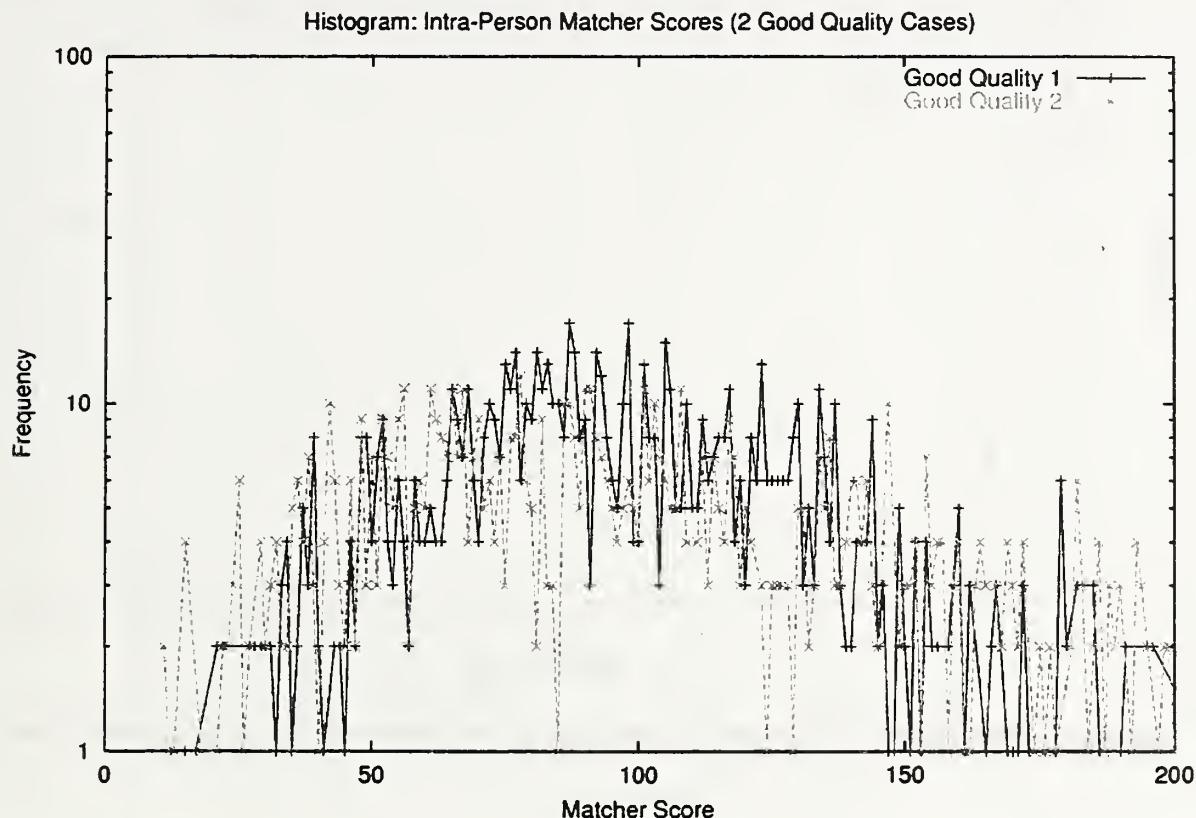


Figure 38. Comparison of Intra-Person Matcher Scores Between Two Good Quality Cases

As an example of non-match distribution, the fingerprints of the two good quality cases can be matched to each other. These scores are easily extracted from the appropriate off-diagonal block in the $3K \times 3K$ similarity matrix.

The distribution of inter-person matcher scores from matching the two good quality cases to each other is plotted in Figure 39. Notice the consistently low magnitude of scores, the mean of which is 8 and the maximum matcher score is 19. These low scores should be expected since the fingerprints being compared are of different classification (whorl and right loop.)

Notice also the small variance on this distribution. The fact that this inter-person distribution has such a small mean and such a low variance indicates that there is very little chance that the identities of these two people will ever be confused based on matching their fingerprints with each other.

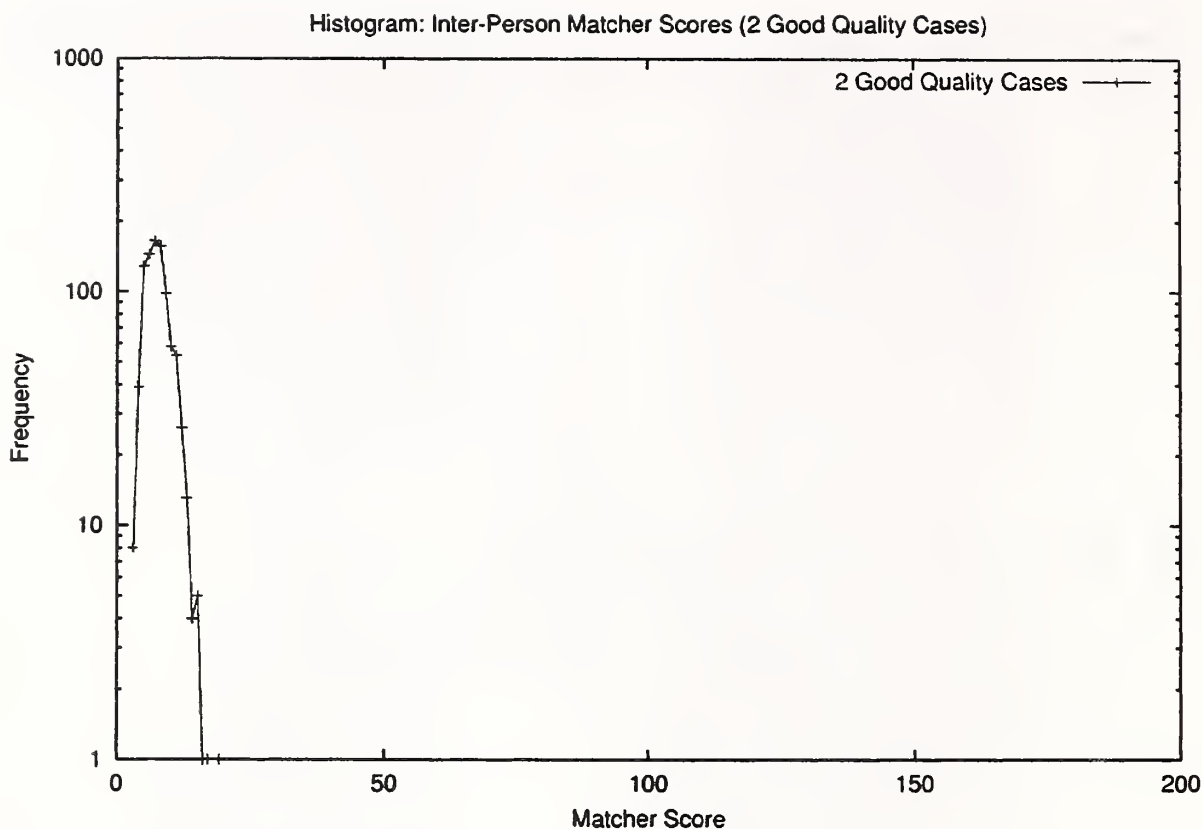


Figure 39. Distribution of Inter-Person Matcher Scores Between Two Good Quality Cases

After studying both intra-person and inter-person scores at the block level, a study was designed to analyze the results based on the entire $3K \times 3K$ similarity matrix. The statistical mean and standard deviation were computed for each of the 10,000 30×30 blocks. The statistics were then separated into two sets, those from intra-person comparison blocks and those from inter-person comparison blocks. Note there are many more inter-person blocks (9,900 off-diagonal blocks) than intra-person blocks (100 main diagonal blocks). The Bozorth98 matcher is essentially symmetric,** so only half of the off-diagonal blocks (those above the main diagonal) were used, totaling 4950 inter-person blocks.

The distribution of means for both of these sets is plotted in Figure 40. Due to the significantly larger number of inter-person blocks than intra-person blocks, frequencies were plotted on the y-axis in log scale.

** A matcher is symmetric if comparing A to B results in the same score as comparing B to A.

Comparing the two mean distributions, notice that the inter-person block means are relatively small with a low variance, while the intra-person block means are mostly larger, but with a relatively large variance.

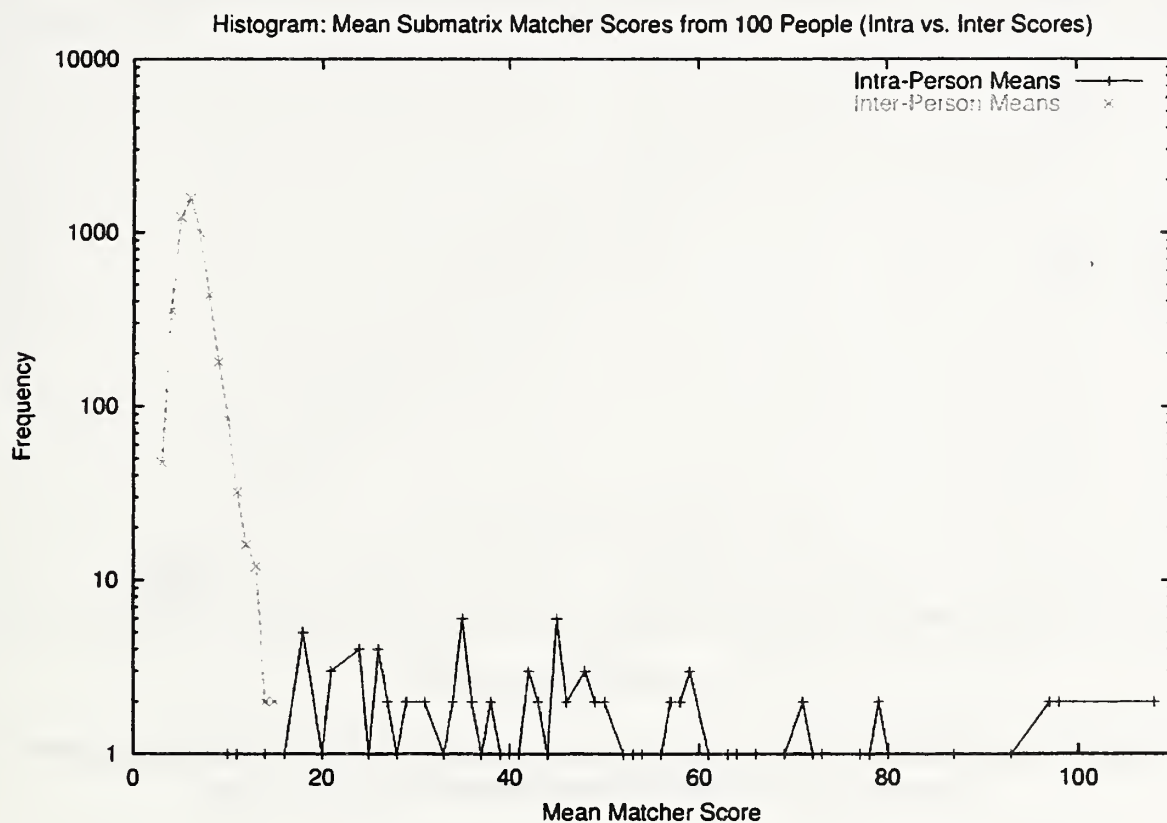


Figure 40. Mean Submatrix Matcher Scores – Comparing intra-person to inter-person submatrices

The distribution of standard deviations for both of these sets is plotted in Figure 41. Notice that the distributions of standard deviations follow a similar pattern to that of the means. Inter-person block standard deviations are relatively small with a low variance, while the intra-person block means are larger, but with a larger variance.

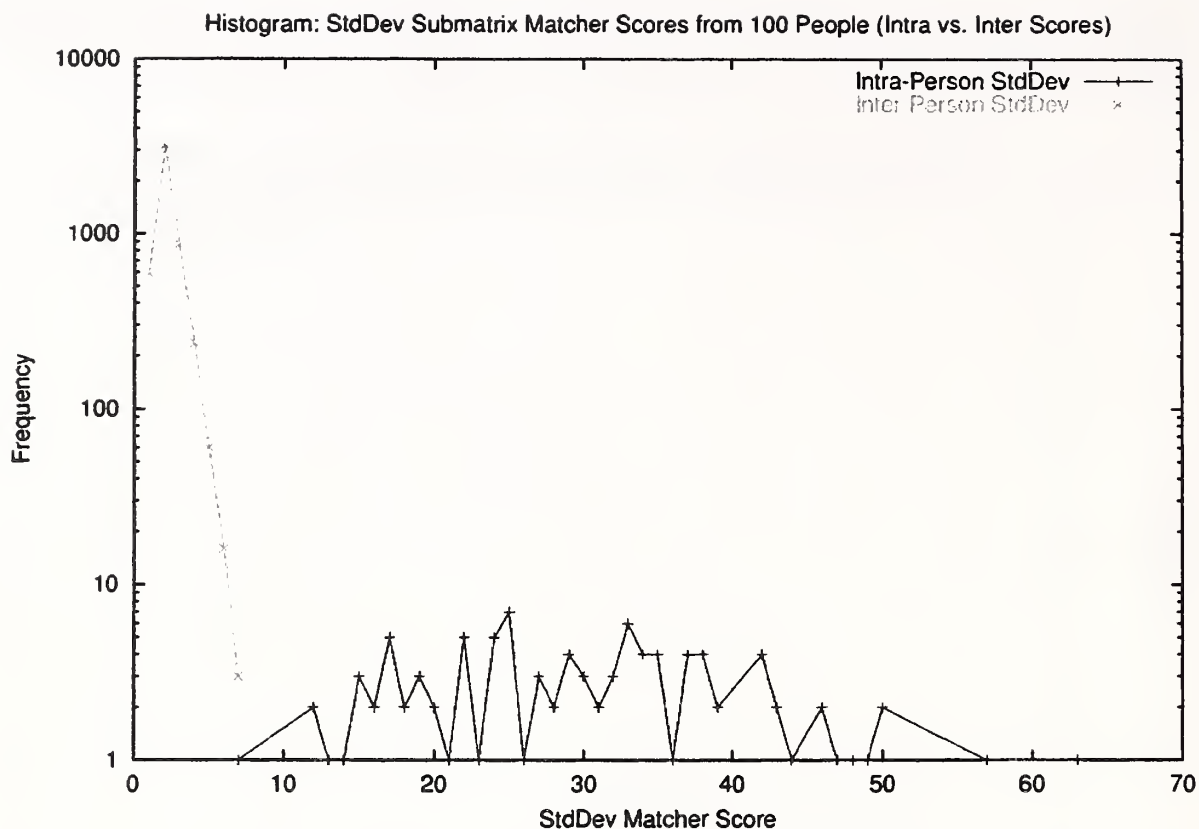


Figure 41. Standard Deviation of Submatrix Matcher Scores – Comparing intra-person to inter-person submatrices

This consistency is further seen within the correlations plot of Figure 42, where the mean of each block is plotted against its corresponding standard deviation. Notice the points are somewhat clustered along a line from the origin of slope 1, and that the inter-person results are clustered near the origin. Due to the high correlation, block analyses involving means should be sufficient.



Figure 42. Correlation Plot of Mean vs. Standard Deviation from Submatrix Matcher Scores - Comparing intra-person to inter-person submatrices

The graph in Figure 40 represents a method for analytically assessing the quality of a fingerprint repository. The lower the intra-person block mean, the more difficult that particular person is to reliably match. The amount of overlap between the mean distributions of intra-person and inter-person block statistics represent how much potential confusability / difficulty exists in a particular repository. More studies need to be conducted to be able to predict a level of performance from a measured amount of overlap, but this technique may be used to empirically compare different repositories. It should be noted that this technique requires there exist multiple (in this case 30) impressions per finger per person in the repository to conduct the analysis.

6. Implications of Metadata

Inevitably with a large scale system performance study, involving the computation and comparison of aggregate statistics such as ROC curves or rank-1 analyses, there follows the question of, "Why?" Without sufficiently organized and labeled data, and without tools to effectively store and access this labeled data, it is very difficult and tedious to analyze underlying events that contribute to an observed result.

Many factors play a potentially complex role in the outcome of a test, or more importantly, in the performance of any operational system. Investigative studies conducted at NIST on the VTB consistently point to image quality as the single most significant factor affecting system performance. Relevant information that may help explain variations in image quality as well as other factors limiting performance may include:

- Capture date
- Capture device
- Capture location
- Operator
- Demographics of the population captured – age, sex, occupation
- Motivation of the population captured - civilian vs. criminal; voluntary vs. compulsory
- Application performed – verification vs. identification

These factors and many more, referred to as *metadata*, may help identify factors that affect or limit performance, and therefore they are important to answering the question, “Why?”

6.1 DHS2 Metadata Study

6.1.1 Nonstationary Results Observed

The results in Section 5.6 shown in Figure 13 were not the first sets of results computed on the DHS2 repository. The $3K \times 3K$ results above were computed on randomly selected subsets of their parent $6K \times 6K$ similarity matrices. Originally, the $6K \times 6K$ matrices were divided in half, based and analyzed using a numerical sort of the person’s identification number (a number sequentially assigned to people based on a complex process of data preparation at NIST.) Lower ordered identification numbers were grouped into the first-half $3K \times 3K$ matrices, while the higher ordered identification numbers were grouped into the second-half $3K \times 3K$ matrices.

Using this non-random criterion for subdividing the parent similarity matrices generated the unexpected results shown in Figure 43. The gray curve (All 6000) in the middle is the $6K \times 6K$ Multi-Trial ROC. The blue curve (First 3000) is the $3K \times 3K$ Multi-Trial ROC corresponding to the lower ordered identification numbers. The green curve (Second 3000) is the $3K \times 3K$ Multi-Trial ROC corresponding to the higher ordered identification numbers. Notice the significant separation in performance between the “First 3000” results and the “Second 3000”, which is quite different from the overlapping results seen in Figure 13.

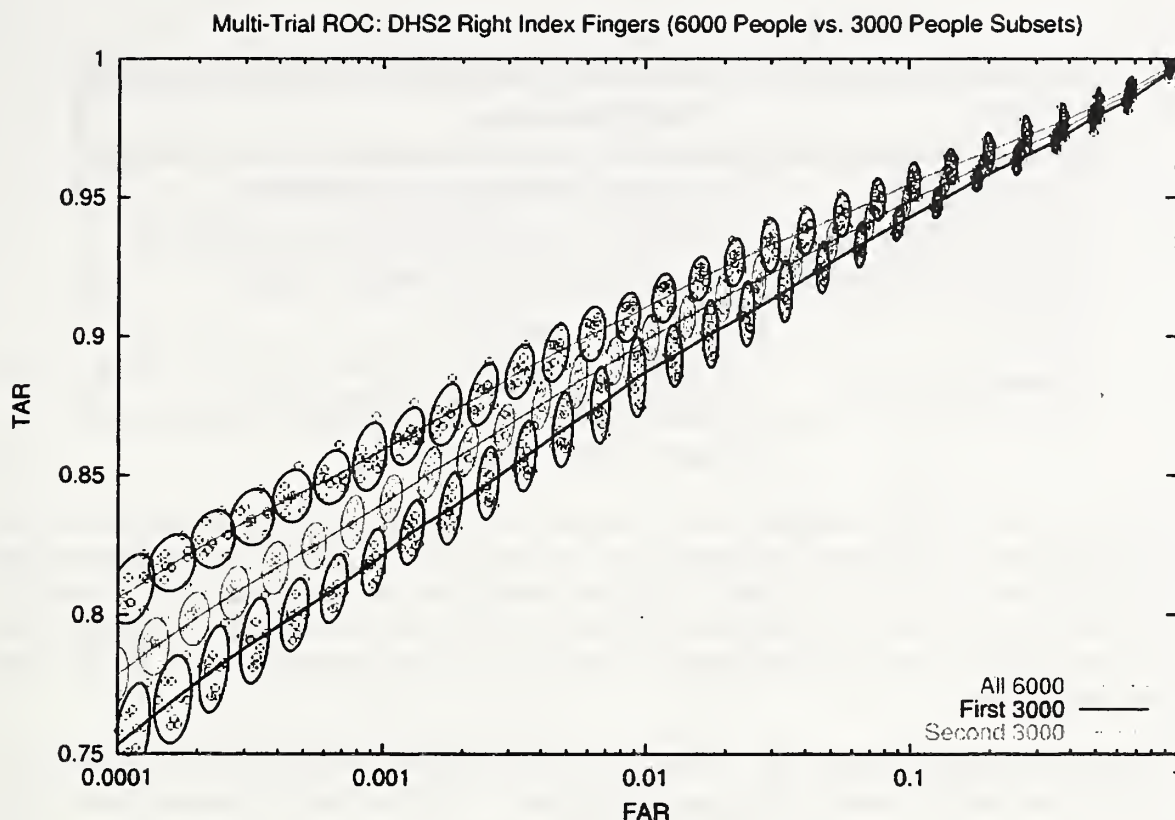


Figure 43. DHS2 Right Index Finger Large Scale Verification – *Procedurally-Selected* Submatrix variation of Multi-Trial ROC

From these results, it is very evident that there is separation in performance between the first and second subsets, and it is highly suspected that the quality of the fingerprints images between the first 3000 people and the second 3000 people are nonstationary. Without further analysis and looking for correlation within associated metadata, not much more can be said.

6.1.2 Metadata Analyzed

Given the separation in performance between the two subsets of ROC results observed in Figure 43, an analysis was conducted on the metadata associated with the cases used in this verification study. It was anticipated that an explanation for the difference in performance could be determined by analyzing the following factors recorded with each fingerprint in the DHS2 repository. (It was anticipated that other observations, important to DHS, would also be made as these factors are analyzed.)

DHS2 Metadata	
Encounter Date	Date when fingerprints were captured
Capture Location	Location of where fingerprints were captured
Cogent Image Quality	Cogent Image Quality Measure (IQM)
Gender	Gender of the subject
Transaction Type	One of 4 types: Asylum, Border Patrol, Inspection, and Border Crossing Card

Table 15. DHS2 Metadata

The large scale DHS2 verification study involved 60K people randomly selected across the entire repository. Each person contributed a pair of right index fingerprints, one fingerprint used as a probe image and the other used as a gallery image; therefore, the study utilized a total of 120K fingerprints. Due to this large sample size, statistics and observations derived from the metadata associated with the fingerprints in this study should be representative of the entire DHS2 repository.

The first factor analyzed was *time*. In the DHS2 repository, a calendar date is recorded when a person's right and left index fingers are live-scanned. Therefore there is a capture date (called the "Encounter Date") recorded for every probe and gallery fingerprint used in this study. These dates are in the form of (mm/dd/yyyy) and were converted and binned into a sequence of quarter indices (called "Encounter Quarters") where the earliest quarter (Quarter 1) is associated with the first quarter of 1995; Quarter 4 is associated with the 4th quarter of 1995; Quarter 5 is associated with the 1st quarter of 1996, and so on and so forth.

The distribution of Encounter Quarters associated with the fingerprints in this study is shown in Figure 44. Note the ramp up of activity from Quarter 1 through Quarter 20. Also, notice the pronounced increase in activity in the fourth quarter of each year.

The red curve in the figure charts the entire set of 60K fingerprint pairs. The blue curve charts the Encounter Quarters associated with the first-half of the fingerprint pairs, while the green curve charts the second-half. These three curves correspond to the three sets of results reported in Figure 43.

As can be seen from the distributions in Figure 44, the majority of cases in the study fall within Quarters 16, 20, and 24. The amounts in which the first-half and second-half cases are represented in these three quarters differ significantly. The first-half cases are heavily represented in Quarter 20, while the second-half cases are more represented in Quarters 16 & 24. It is logical to hypothesize that the difference in performance between the first and second-half cases can be attributed to characteristic differences between these three quarters.

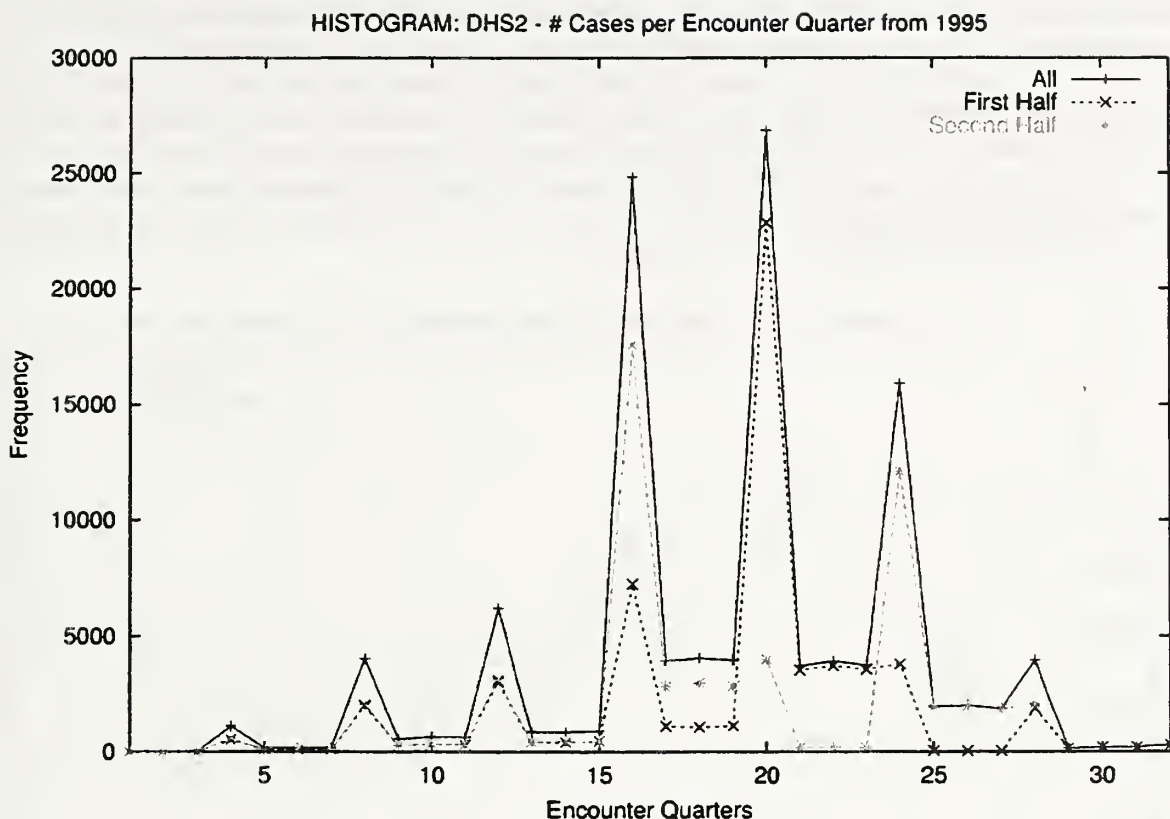


Figure 44. Frequency of Encounters across Time

An image quality measure (IQM) is also recorded with each fingerprint in the DHS2 repository. This is a proprietary quality measure developed by Cogent. IQM values are computed on a scale of 1 to 8, with 1 being the highest quality and 8 being the lowest. It was determined to be interesting to analyze how quality (expressed in terms of IQM) might vary over time in the repository.

Little is publicly known about the Cogent IQM algorithm. A mean IQM value may be used to represent the quality within a sample of fingerprints, but as there are a limited number of (eight) levels reported, and the distribution characteristics across the potential range of these levels is unknown, a different method was tested and applied. With 1 being the highest quality level, it is logical when capturing fingerprint images to desire to have as many fingerprints as possible with an IQM value of 1; therefore, the percentage of fingerprints with IQM equal to 1 was computed and used to represent how “good” a repository (or subset) is.

Figure 45 plots this measure of “good” quality across each quarter of the DHS2 repository. Notice an overall (red curve) decline in quality from Quarter 9 through Quarter 23. This demonstrates that fingerprint image quality is nonstationary. Up till Quarter 24, the second-half

cases (green curve) are of consistently better quality than first-half cases (blue curve). This observation supports the second-half cases performing better than the first-half.

The disparity between first and second half qualities is even greater when examining the high-activity Quarters 16, 20, & 24. In fact, the difference between the two sets within Quarter 20 is one of the largest across the entire series. Quarter 20 is of significantly lower quality and, as seen in Figure 44, primarily comprised of first half cases, which again support the second half cases performing overall better than the first half.

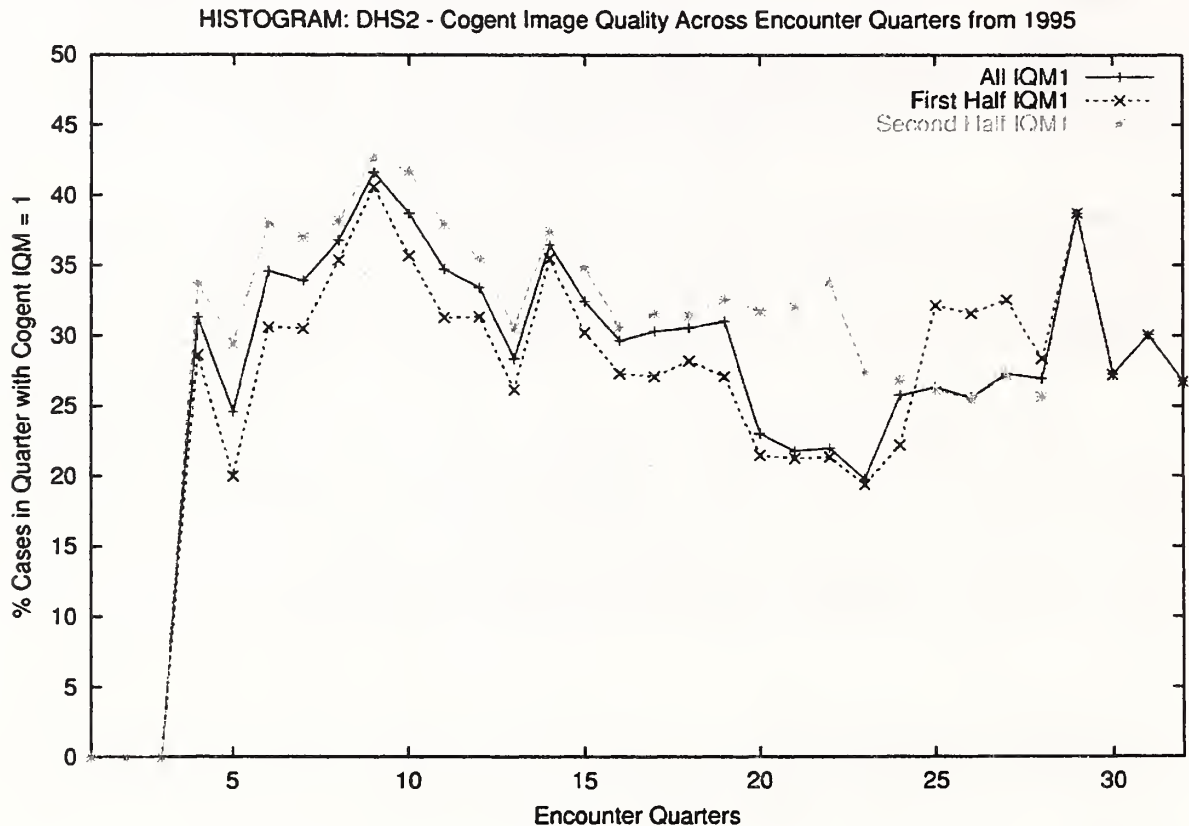


Figure 45. Quality of Images Captured across Time

Figure 46 plots the change in gender across time. The statistic used here is the percentage of female cases within each quarter. Female statistics are used here as opposed to male due to the hypothesis that female fingerprints are generally more difficult to match than male.

Looking at the overall (red) curve, there is a trend of an increasing percentage of female cases, starting with about 5% at Quarter 4 and increasing to 15% at Quarter 20.

Again, when examining differences between first and second-half statistics at Quarters 16, 20, and 24, notice the first-half cases for Quarters 20 has somewhat (4%) more females than the first-half. Recall from Figure 44 that Quarter 20 is mostly comprised of first-half cases, and we

see here that there are more females represented in the first-half than the second half. This lends support to the hypothesis that females are more difficult to match, as the first-half performs worse than the second-half and there is evidence that the ratios of gender do differ.

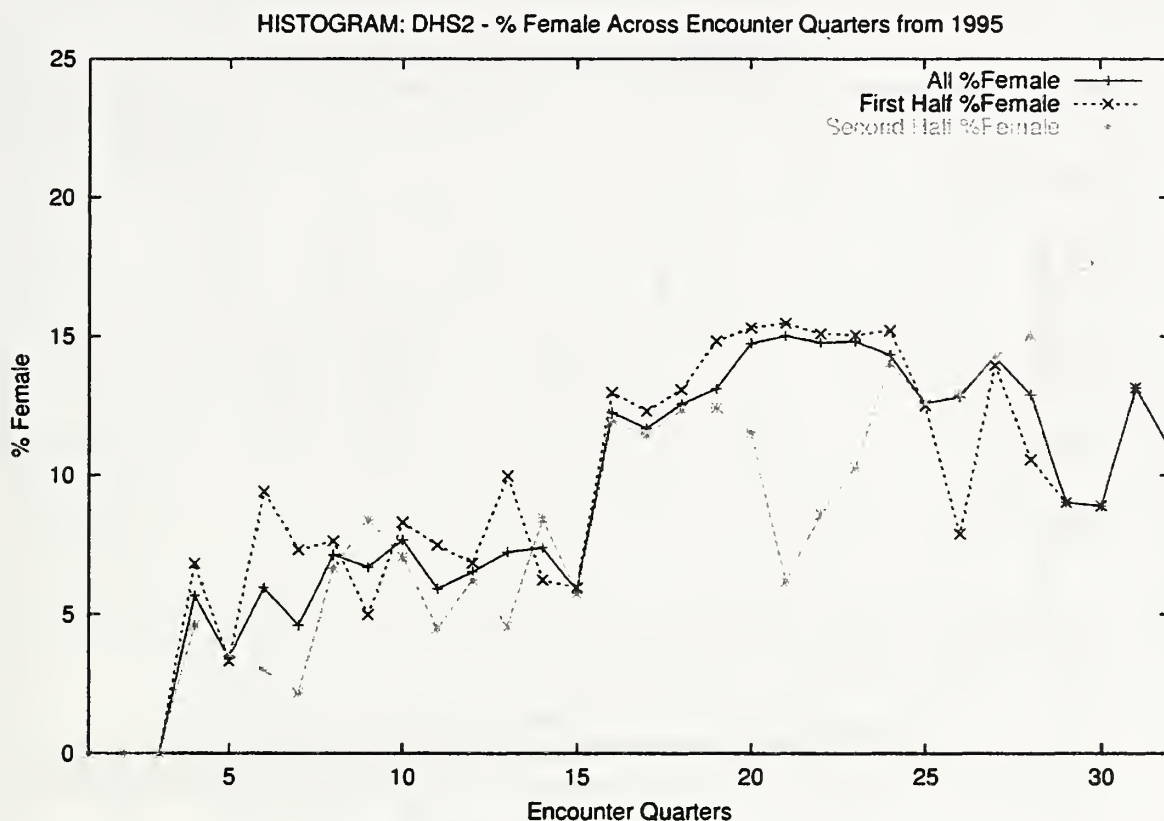


Figure 46. Gender across Time

Another attribute recorded for each case in DHS2 is the location at which the fingerprints were captured (called the "Capture Location"). There are over 300 different Capture Locations recorded in the 60K cases of this study. Figure 47 plots a histogram of the 50 most frequent locations, representing 88% of the cases in this study; 48% of the cases are represented in the first 10 locations; 33% in the first 5 locations.

Figure 48 plots image quality (% of cases with IQM = 1) across the 50 most frequent Capture Locations. Notice the wide variation in quality across locations and the exceptionally high quality with locations such as Capture Location 6. In general, the quality of second-half (green curve) cases is higher than first-half (blue curve) cases at nearly every location.

The next figure, Figure 49, plots gender (% female cases) across Capture Locations. As was the case with image quality, the percentage of females significantly fluctuates from location to location. Of the more frequent locations, the most dramatic increase is with Capture Location 6 where overall (red curve) 40% of the cases at this location are female. A curious observation is

that Capture Location 6, although it has a high percentage of female cases, has significantly high image quality according to Figure 48. Also notice in Figure 49 that with a majority of locations, the first-half (blue curve) cases have a higher percentage of females than the second-half (green curve).

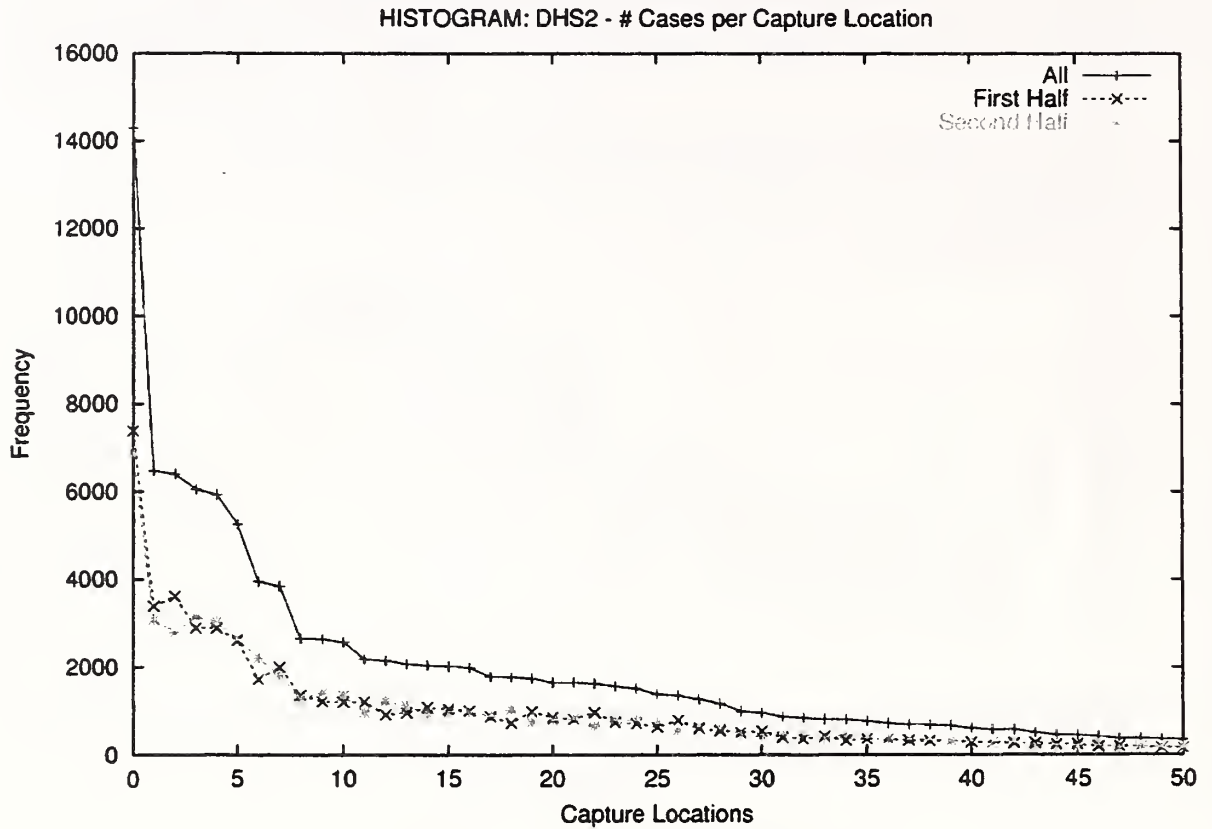


Figure 47. Frequency of Encounters across Capture Location

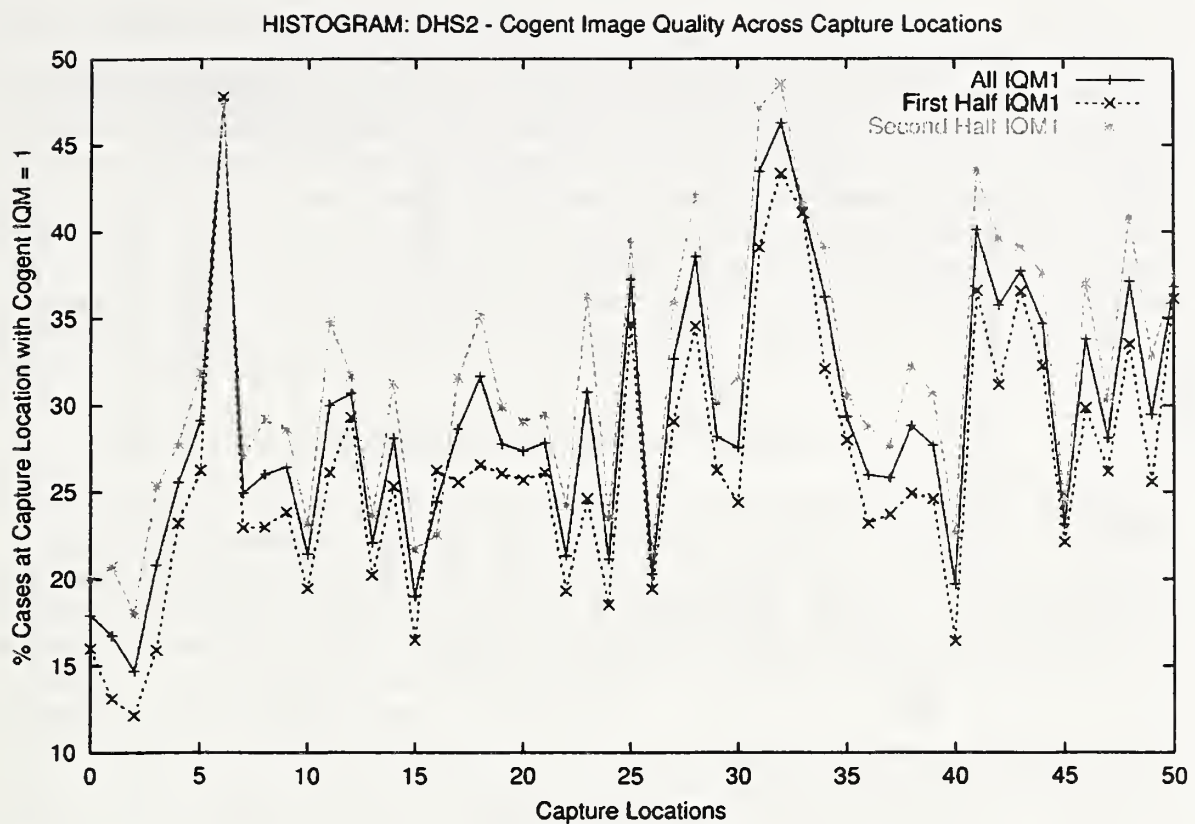


Figure 48. Quality of Images across Capture Location

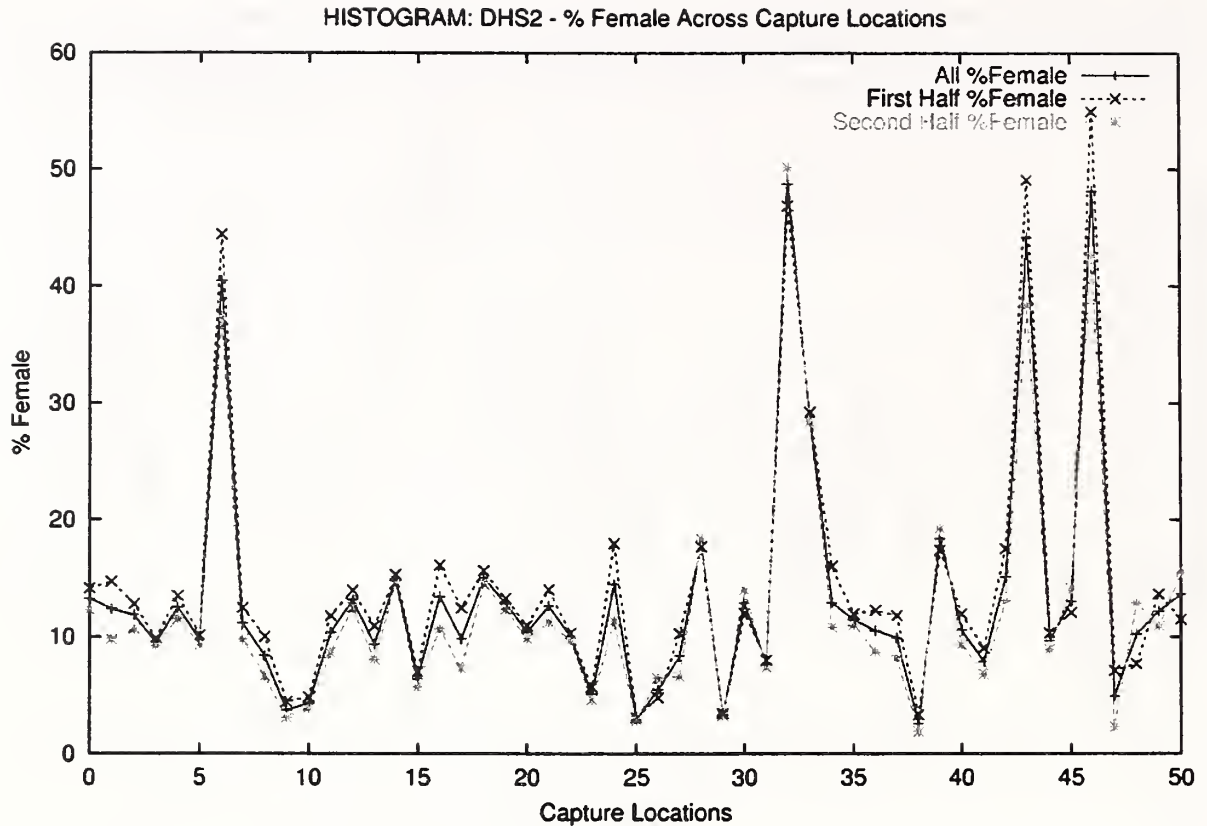


Figure 49. Gender across Capture Location

While some interesting and perhaps useful qualitative observations can be made from the metadata plots above, it is difficult (if not impossible) to derive quantitative measurements from these graphs that accurately describe the effect these different metadata factors have on performance. To take a more statistical view of these factors, a correlation matrix was computed between the integer data representing capture quarters, capture location, image quality, gender, and transaction types.

Looking at the correlation coefficients in Table 16, the strongest correlations are between transaction type and gender with a value of 0.29, and between capture quarter and gender with a value of 0.2. Neither of these are indications of strong correlations.

	Quarter	Location	Quality	Gender	Transaction
Quarter	1.000000000	-0.003495314	0.06894202	0.20043526	-0.04144020
Location	-0.003495314	1.000000000	-0.13102489	0.05149515	0.05456955
Quality	0.068942024	-0.131024892	1.000000000	0.01373389	-0.02282792
Gender	0.200435257	0.051495149	0.01373389	1.000000000	0.28768644
Transaction	-0.041440196	0.054569546	-0.02282792	0.28768644	1.000000000

Table 16. Correlation Table of Metadata Factors

6.1.3 DHS2 Metadata Study Summary

By analyzing the metadata recorded with DHS2, several important discoveries were made. First, we observed that both image quality and gender were nonstationary, as they varied significantly over time and across different capture locations. Second, aggregate ROC analyses may hide these types of nonstationary factors, which are important to understanding and improving the performance of operational systems. This was demonstrated by comparing the separated results of Figure 43 (where subsets were selected procedurally based on an identification number) to the overlapped results of Figure 13 (where subsets were selected randomly). This supports the third conclusion, that biometric performance studies should include tests for discovering and analyzing nonstationary factors.

6.2 Concept for a Fingerprint Experiment Manager

As described above, DHS2 repository has an extensive set of metadata associated with each pair of fingerprints captured, while the HEF framework utilized to generate performance statistics in this report provides virtual probe and gallery sets to be defined from within a large similarity matrix. With the data and framework in place, NIST has the ability to conduct experiments where probe and gallery sets of fingerprints are dynamically and strategically selected. By computing performance statistics on different probe and gallery sets, specific factors can be isolated and their effect measured.

A challenge however exists. As the size of a repository grows, the ability to manually select, archive, and compare results from multiple experiments becomes prohibitive. Even more, specific pairs of probe gallery fingerprints matched in a prior experiment may be selected again for a new experiment given new criteria. In this case, it would be advantageous to retrieve the previous match score rather than compute a new one.

To automate all this bookkeeping, a database-driven experiment manager is needed. This is the topic of ongoing research and development at NIST. Conceptually, this database tool is initially loaded with a repository of fingerprints and associated metadata. Metadata queries are constructed based on factors of interest and ran against the database returning a set of relevant probe and gallery fingerprints to be matched. Those matches not previously computed between the probe and gallery sets are computed and stored in a repository-level similarity matrix. After which performance statistics are computed and reports generated. Then a new metadata query is constructed and a new set of experiments ran.

A tool such as this holds great potential in gaining understanding of the performance of a fingerprint matching system.

7. Conclusions

This report documents a large set of evaluations performed on diverse data sets using a fingerprint verification system developed at NIST in cooperation with the FBI. This system was designed to allow two functions to be performed. First the system serves as an open system for

the evaluation of fingerprint technology which allows different datasets and evaluation methods to be tested. Second, the system sets minimum standards of performance for fingerprint systems. The software used to build the system is publicly available at no cost so any system that can't out-perform the VTB in either accuracy or speed should also have very minimal cost.

7.1 Critical Test Parameters

The conclusions presented here which involve datasets have used two distinct types of data. The first type of data includes the NIST Special Databases such as SD14, SD24, and SD29. These are small data collections (200-2000 subjects) that contain only images that are good enough to be matched by an AFIS or by human examiners or both. Given the methods by which these small datasets were collected, all failures to acquire and failures to enroll have been eliminated. In other words, all cases were removed where the captured fingerprint image was of sufficiently poor quality so that using it to match would be impossible. These datasets have been widely used for research and development of fingerprint systems and reflect the type of data that might be available using well controlled laboratory conditions. Comparison of algorithms using this data can be duplicated by other workers in the field and should produce results similar to previously published studies.

The second type of data is large datasets which were collected with no prior plan to use the fingerprints for evaluations. The usual collection procedure for this data is to try to collect the best fingerprints possible in a fixed but limited time. If none of the fingerprints collected are as good as could be obtained with an unlimited time, you submit the best set available. These datasets contain measurable numbers of fingerprints that are not usable and that would be failures to acquire or failures to enroll. Examples of this type of data are DHS2, DHS10, and TXDPS. In these datasets the unusable fingerprints lower the TAR by some fixed amount. In Section 5.13 this number for the DHS2 dataset was measured as 2%. Additional studies using commercial AFIS system are underway to evaluate these rates for other datasets and systems.

In all the comparisons of results discussed here, three factors need to be considered. First, does the data contain the type of information needed to calculate the failure to acquire rate? Second, is the data stationary over time? This not only implies consistent data collection procedures but uniform demographics. Third, is the sample variance known or is the sample size large enough to allow confidence limits to be measured? In the NIST Special Databases the images that would cause acquisition failures have been removed. In the DHS, TXDPS, and DOS datasets, images that are of poor enough quality to cause acquisition failures are retained. All of the large datasets used here were tested for stationary statistics and only DHS2 was nonstationary. DHS2 is the only dataset which was collected at locations where the demographics have been observed to be time dependent. The two standard deviation variance of the SD sets was not calculated because of small sample sizes. The two standard deviation variance in verification rate of the large samples ranges from 1% for DHS2, to 5% for TXDPS, with DHS10 and DOS data at about 2%; since DHS2 is nonstationary, its variance is dependent on sampling.

7.2 Small Sample Test Conclusions

For single fingers, plain-to-plain matching is substantially less accurate than rolled-to-rolled matching. This is clearly illustrated by Figure 5 in which the rolled-to-rolled TAR at 1% FAR is

98% while the plain-to-plain TAR is 95%. It is common for the FRR ($1 - \text{TAR}$) to be two to three times higher for plain-to-plain matching than for rolled-to-rolled matching. As an example consider the data shown in Figure 6. The TAR at 1% FAR for rolled-to-rolled fingerprints is 96% and for plain-to-rolled matching is 90%. Plain-to-rolled matching is usually somewhat less accurate than plain-to-plain matching. In Figure 5 through Figure 9, the plain-to-rolled ROC curve is near or below the plain-to-plain ROC in all cases. For Figure 8, the ring finger, plain-to-rolled matching is substantially below plain-to-plain matching. Preliminary results for commercial systems confirm that plain-to-rolled matching is substantially less accurate than rolled-to-rolled matching. Further tests on plain-to-plain matching are being conducted on commercial systems.

Since matcher scores for different fingers are nearly statistically independent, combining two fingers using a linear 2-D threshold is very effective. Figure 22 and Figure 24 illustrate that matcher scores for thumbs and index fingers respectively are uncorrelated. These figures also show that a simple linear discriminate with a slope of -1 will serve as a simple boundary between match and nonmatch scores. When index finger and thumb scores are combined using this discriminate, it is possible to produce verification results with a TAR of 99% at a FAR of 1%. Combining right and left index finger matcher scores in the same way is somewhat less effective but still allows a TAR of 98% to be achieved at a FAR of 1%.

For the matching of plain fingerprints, area is very important. Thumbs are more effective than index fingers and little fingers are of very limited value. This is clearly illustrated for live-scan data in Figure 4 and for plain fingerprint images taken from inked card in Figure 5 through Figure 9. Most commercial AFIS systems make extensive use of index finger pairs for rolled-to-rolled matching. This strategy was adopted in identification applications because studies have shown that index fingers are effective when used for pattern classification-based filtering, particularly when compared to other fingers. This does not appear to be a good strategy for verification using plain impressions. Finger area is more important and no filtering is required. The larger area of thumbs makes them a better candidate for verification.

7.3 Large Sample Test Conclusions

Tests on large datasets show similar single finger matching results and sample-related variations to those seen with the small datasets. There is however a significant variation in sample results that is related to image quality and the variation of image quality about the mean value. These results can not be adequately tested on small datasets since the small samples used were selected after matching and human visual inspection were conducted, which produces images with greater quality than one would expect to find in unscreened operational data.

SD29 index finger plain-to-plain matching results on the VTB are similar to DHS2 and DOS plain-to-plain index finger results as shown in Figure 50. This is not the result that was expected. SD29 represents 20 year-old good quality inked data that was scanned and checked at NIST. DHS2 is data collected by the former INS under field conditions using live-scan equipment. The TAR at 1% FAR for SD29 is 88% while the TAR at 1% FAR for DHS2 and DOS is 90%. The TAR at 0.01% FAR for SD29 is 76% while the TAR at 0.01% FAR for DHS2 is 78%, and for DOS it is 79%. This also shows that results for operational tests using live-scan data (DHS2 or DOS) are statistically the same as results using screened inked fingerprints

(SD29). The expected loss of quality going from good quality inked cards to operational quality fingerprints is more than compensated for by the improvement in live-scan image quality.

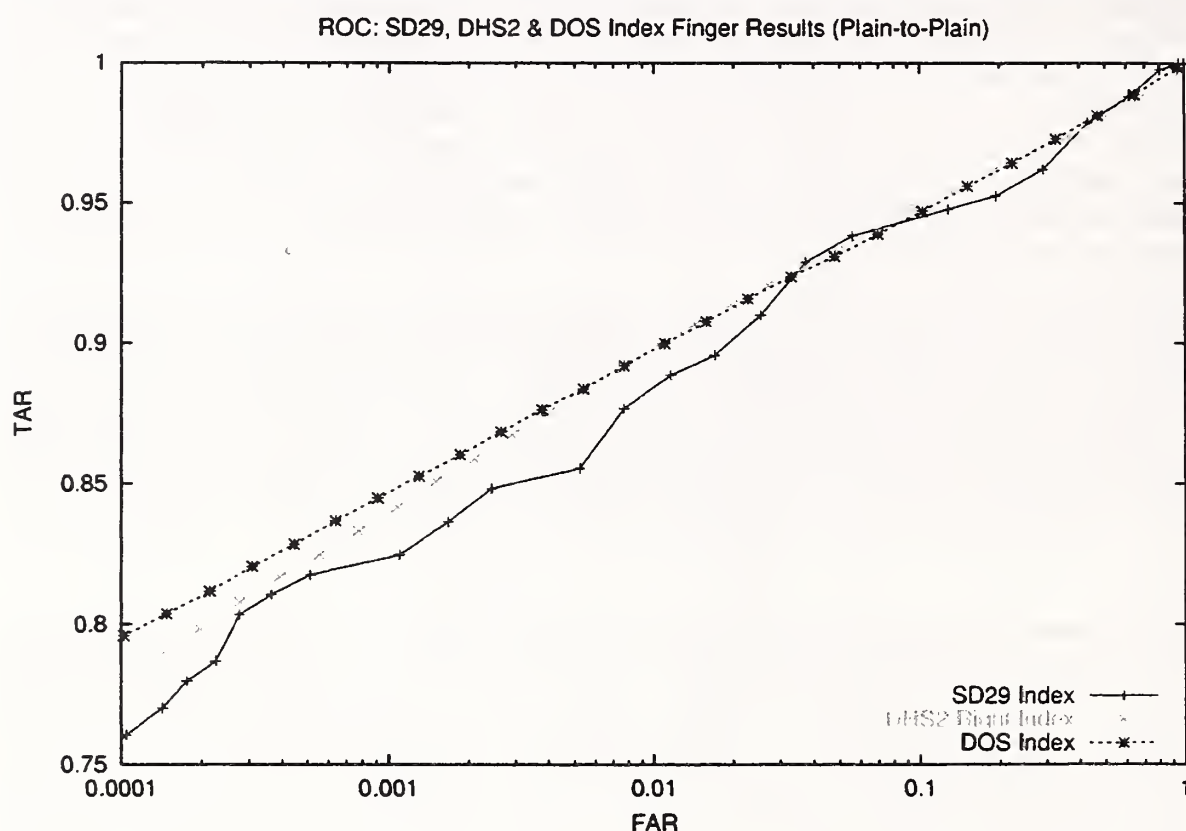


Figure 50. Comparison of Plain-to-Plain Results from SD29, DHS2, & DOS

As shown in Figure 51, SD24 index finger plain-to-plain matching results are similar to TXDPS plain-to-rolled index finger results. SD24 data was collected under laboratory conditions at NIST using live-scan equipment. TXDPS data is a mixture of inked and live-scan data collected under widely varying conditions. (These conditions cause a large variance in the TXDPS data.) The TAR at 1% FAR for SD24 is 93% while the TAR at 1% FAR for TXDPS is 92%. The TAR at 0.01% FAR for SD24 is 83% while the TAR at 0.01% FAR for TXDPS is 85%. The lower bound on the TXDPS data results is a TAR of 88% at a FAR of 1% and a TAR of 81% at a FAR of 0.01%. This is better than either SD29 or DHS2 and demonstrates that current law enforcement data can be collected with image quality approaching laboratory quality results. It further suggests that improved collection or sensor technology should be able to bring the results of large scale data collections such as DHS2 and DOS close to the results of SD24 and TXDPS. At the same time Figure 51 shows the ROC results for DHS10 data. This data yields substantially lower TAR for any given FAR and is an indication of the type of results that might be expected with some archival data set in existing systems.

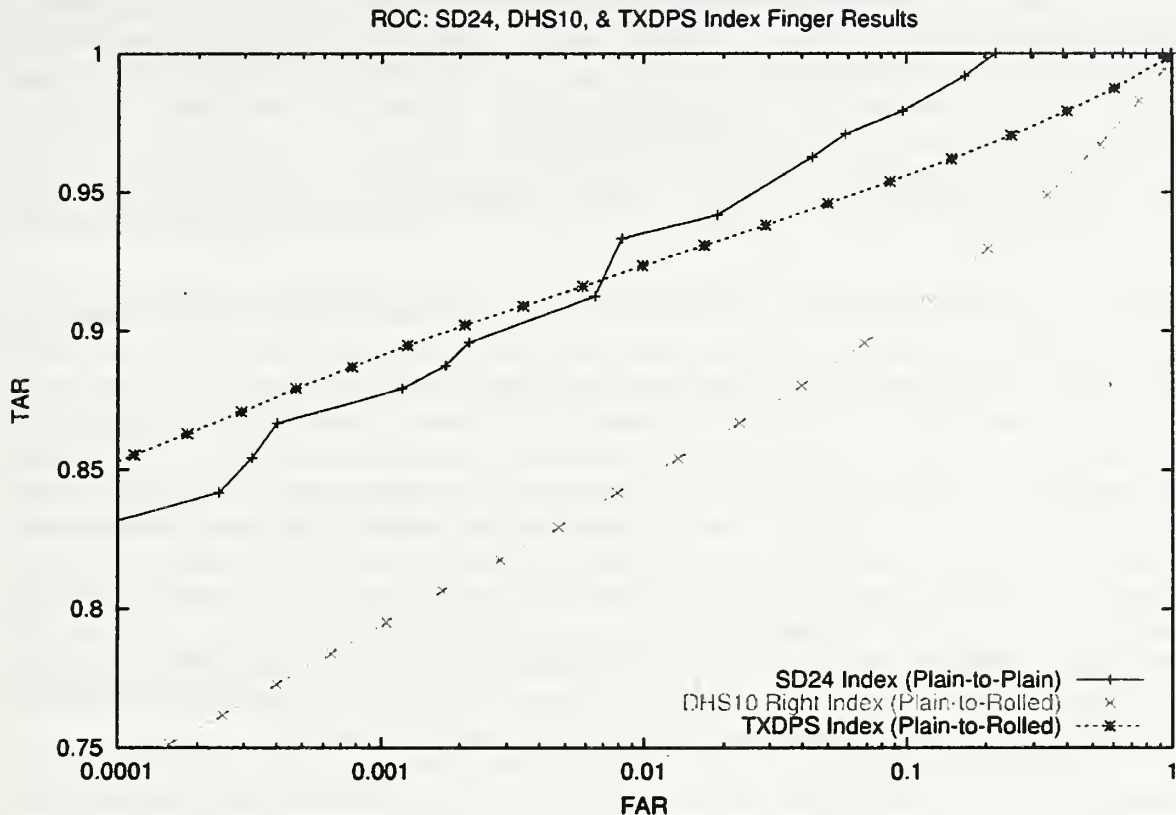


Figure 51. Comparison of Results from SD24, DHS10, & TXDPS

The splitting of DHS2 verification results shown in Figure 43 demonstrates that large samples must be checked for stationary properties. Analyzing the metadata recorded with DHS2, several important discoveries were made. First, we observed that both image quality and gender were nonstationary, as they varied significantly over time and across different capture locations. Second, aggregate ROC analyses may hide these types of nonstationary factors, which are important to understanding and improving the performance of operational systems. This was demonstrated by comparing the separated results of Figure 43 (where subsets were selected procedurally based on an identification number) to the overlapped results of Figure 13 (where subsets were selected randomly). This supports the third conclusion, that biometric performance studies should include tests for discovering and analyzing nonstationary factors. The ability of studies of the type discussed here to predict deployed system performance will depend on how well the correlation between demographic and location-dependent factors and biometric performance are understood. Treating only the average performance derived from test samples that do not reflect these factors will produce incorrect results. Fingerprint verification performance can not be treated as a single number that covers all systems or all demographic mixes.

The identification results shown in Figure 18 demonstrate as much as a 10% fluctuation in correct identification across ten, 100-finger rank-1 identification trials using a 620,000 fingerprint gallery from DHS2 data. The same data also shows that it is possible for the average results in 100-finger samples to be 10% different between right and left index fingers. Figure 19 shows the average results for this experiment as a function of gallery size. Above a gallery size of 10, the right index finger consistently has a higher identification rate than the left index finger. The VTB's identification rate for a gallery size of 620,000 is 76% for right index fingers, using rank-1 thresholding.

Since most commercial AFIS systems use two or more fingers successfully for identification and since the NIST recommendation to Congress is for two or more fingers, limited tests were performed for identification using more than one finger. Combining two index fingers using rank-based fusion improves the 620,000 fingerprint identification results by 9%. The method used to combine fingers is one which produces minimal computational load. The left index fingers of the top-100 right index fingers are tested and the scores are summed and ranked. The effect of this combined scoring is shown in Figure 28. The details of this process are discussed in Section 5.12.3. It should be kept in mind that this procedure requires that the scores of the two fingers used be largely statistically independent. If a match is missed because of poor image quality and a second finger image of equally poor quality is used, the combined score may still not produce a satisfactory match. This effect is illustrated in Figure 31. Using 1000 fingerprints 1.8% could not be scored as rank-1 and 0.2% fell below rank-1 using this combining procedure, so that 2.0% of the identifications were not improved by combining scores.

Sections 5.12.1 & 5.12.2 discuss a very simple form of score-based fusion using two fingers for identification. The results of these studies are shown in Figure 23 through Figure 27. While the fusion is less than ideal as shown in tables 10 and 12, the results make a dramatic improvement in the single finger verification rates. These results are the basis for the NIST recommendation for two or more fingerprints in Reference (303a) [34]. Using a thumb and index finger from SD29, a TAR of 98.6% can be achieved at a FAR of 1%. Using two index fingers from SD29 a TAR of 97.6% can be achieved at a FAR of 1%.

Since 2.0% of the identifications in the two-finger identification experiment failed to improve the results of a single-finger match, and since no data exists on the repeatability of large scale fingerprint samples, a study of fingerprint repeatability was done using DHS2 data. Fingerprint matching repeatability results show that, for 98% of subjects, fingerprint matching is highly repeatable. For the other 2% of subjects, fingerprints of adequate quality for matching can not be obtained. This test is important for two reasons. First it allows the repeatability and image quality to be evaluated solely on the basis of matcher performance. Second, since sample size for each individual was 100, it allowed visual inspection of the image sets that were not repeatable. This visual inspection confirmed that the non-repeatable image usually appeared to be of adequate contrast and resolution but that the friction ridges were not well defined.

This document has introduced the VTB and results of several studies made possible by the VTB. The VTB will have an ongoing role as a platform for a variety of fingerprint studies in the future, building on these results. As part of this ongoing work, an appendix comparing the VTB matcher to two commercial fingerprint systems has been added to this report. This appendix

concludes that the performance of the VTB is very similar to commercial verification systems currently on the market.

REFERENCES

- [1] J.H. Wegstein, "A Semi-automated Single Fingerprint Identification System," NBS Technical Note 481, April 1969.
- [2] J.H. Wegstein, "Automated Fingerprint Identification," NBS Technical Note 538, August 1970.
- [3] J.H. Wegstein, "Manual and Computerized Footprint Identification," NBS Technical Note 712, February 1972.
- [4] R.T. Moore, "The Influence of Ink on The Quality of Fingerprint Impressions," NBS Technical Report NBSIR 74-627, December 1974.
- [5] J.H. Wegstein, "The M40 Fingerprint Matcher," NBS Technical Note 878, July 1975.
- [6] J.H. Wegstein, and J.F. Rafferty, "The LX39 latent Fingerprint Matcher," NBS Special Publication 500-36, August 1978.
- [7] R.T. Moore, "Results of Fingerprint Image Quality Experiments," NBS Technical Report NBSIR 81-2298, June 1981.
- [8] J.H. Wegstein, "An Automated Fingerprint Identification System," NBS Special Publication 500-89, February 1982.
- [9] R.M. McCabe, and R.T. Moore, "Data Format for Information Interchange," American National Standard ANSI/NBS-ICST 1-1986, August 1986.
- [10] R.T. Moore, "Automated Fingerprint Identification Systems - Benchmark Test of Relative Performance," American National Standard ANSI/IAI 1-1988, February 1988.
- [11] R.T. Moore, "Automated Fingerprint Identification Systems – Glossary of Terms and Acronyms," American National Standard ANSI/IAI 2-1988, July 1988.
- [12] R.T. Moore, R.M. McCabe, and R.A. Wilkinson, "AFRS Performance Evaluation Tests," NBS Technical Report NBSIR 88-3831, August 1988.
- [13] "Minimum Image Quality Requirements for Live Scan, Electronically Produced Fingerprint Cards," Technical Report for the Federal Bureau of Investigation – Identification Division, November 1988.
- [14] C. Watson, "NIST Special Database 4: 8-bit Gray Scale Images of Fingerprint Image Groups." CD-ROM & documentation, March 1992.

- [15] C.L. Wilson, G.T. Candela, P.J. Grother, C.I. Watson, and R.A. Wilkinson, "Massively Parallel Neural Network Fingerprint Classification System," Technical Report NISTIR 4880, July 1992.
- [16] R. McCabe, C. Wilson, and D. Grubb, "Research Considerations Regarding FBI-IAFIS Tasks & Requirements," NIST Technical Report NISTIR 4892, July 1992.
- [17] G.T. Candela and R. Chellappa, "Comparative Performance of Classification Methods for Fingerprints," Technical Report NISTIR 5163, April 1993.
- [18] C. Watson, "NIST Special Database 9: 8-Bit Gray Scale Images of Mated Fingerprint Card Pairs," Vol. 1-5, CD-ROM & documentation, May 1993.
- [19] C. Watson, "NIST Special Database 10: Supplemental Fingerprint Card Data (SFCD) for NIST Special Database 9," CD-ROM & documentation, June 1993.
- [20] C. Watson, "NIST Special Database 14: Mated Fingerprint Card Pairs 2," CD-ROM & documentation, September 1993.
- [21] R.M. McCabe, "Data Format for the Interchange of Fingerprint Information," American National Standard ANSI/NIST-CSL 1-1993, November 1993.
- [22] J.L. Blue, G.T. Candela, P.J. Grother, R. Chellappa, C.L. Wilson, "Evaluation of Pattern Classifiers for Fingerprint and OCR Application," in Pattern Recognition, 27, pp. 485-501, 1994.
- [23] C.L. Wilson, G.T. Candela, C.I. Watson, "Neural Network Fingerprint Classification," in Journal for Artificial Neural Networks, 1(2), 203-228, 1994.
- [24] C.I. Watson, J. Candela, P. Grother, "Comparison of FFT Fingerprint Filtering Methods for Neural Network Classification," Technical Report NISTIR 5493 September 1994.
- [25] C. Watson, "NIST Special Database 18: Mugshot Identification Database of 8 bit gray scale images," CD-ROM & documentation, December 1994.
- [26] G.T. Candela, P.J. Grother, C.I. Watson, R.A. Wilkinson, C.L. Wilson, "PCASYS - A Pattern-level Classification Automation System for Fingerprints," Technical Report NISTIR 5647 & CD-ROM, April 1995.
- [27] R.M. McCabe, "Data Format for the Interchange of Fingerprint, Facial & SMT Information," American National Standard ANSI/NIST-ITL 1a-1997, April 1997.
- [28] C. Watson, "NIST Special Database 24: Digital Video of Live-Scan Fingerprint Data," CD-ROM & documentation, July 1998.
- [29] M.D. Garris and R.M. McCabe, "NIST Special Database 27: Fingerprint Minutiae From Latent and Matching Tenprint Images," CD-ROM & documentation, June 2000.

[30] R.M. McCabe, "Data Format for the Interchange of Fingerprint, Facial, & Scar Mark & Tattoo (SMT) Information," American National Standard ANSI/NIST-ITL 1-2000, July 2000. Available from R.M. McCabe at NIST, 100 Bureau Drive, Stop 8940, Gaithersburg, MD 20899-8940.

[31] D.M. Blackburn, J.M. Bone, and P.J. Phillips, "Facial Recognition Vendor Test (FRVT) 2000 Results," PDF documents at www.frvt.org/FRVT2000/documents.htm, February 2001.

[32] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone, "Face Recognition Vendor Test 2002 Results," PDF documents at www.frvt.org/FRVT2002/documents.htm, March 2003.

[33] R.J. Micheals, P.J. Grother, and P.J. Phillips, "The NIST HumanID Evaluation Framework," PDF document at www.frvt.org/EvalMethod.htm, 2003.

[34] "Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability," PDF document at <http://www.itl.nist.gov/iaui/894.03/fing/fing.html>, November 2002.

[35] M.D. Garriss, C.I. Watson, R.M. McCabe, C.L. Wilson, "NIST Fingerprint Image Software (NFIS)," User's Guide - Technical Report NISTIR 6813 & CD-ROM, October 2001.

APPENDIX A. Matcher Comparisons to the VTB

At the time this document was written no data comparing the VTB with other matchers was available. During the review process, tests that enabled comparison of the VTB matcher with commercial products were started and some data of interest has become available. This data is shown as a set of ROC curves in Figure 52. All of the curves are derived from 6K×6K similarity matrices. The figure contains a total of twelve ROC curves. Three matchers (the VTB matcher and two commercial products) were tested on four sets of data.

In all cases the suppliers were told that the primary data of interest was the data set from the Department of State Mexican visa program (DOS). After this report went to editorial review, a new highest quality data set was obtained from DHS. This data set, referred to as BEN, was acquired using new live-scan equipment and has been subjected to stringent human quality control. Further tests with the BEN data are ongoing.

The data presented in Figure 52 and Table 17 shows that the quality of data is as important for all the algorithms tested as is the particular algorithm being tested. The best results presented are for plain-to-rolled matching of the right thumb on BEN data. The two best algorithms are NIST's VTB and vendor A. For the VTB (blue diamond) and vendor A (red square) a TAR of 98% is achieved at 1% FAR. For the DOS data set using plain-to-plain index fingers at 1% FAR, vendor A (red plus) has a TAR of 94.5%, vendor B (blue square) has a TAR of 96% and the VTB (gray triangle) has a TAR of 91.4%. In the first case (BEN right thumb plain-to-rolled) the VTB tied the best vendor, and in the second (DOS) case both vendors were better than the VTB.

The two worst cases are vendor B for Texas data (yellow circle) and the VTB for DOS data (gray triangle). This indicates that the results presented in this report are worst case for the VTB, but are better than one could expect for commercial software which is not optimized for Texas quality data.

From this limited comparison we conclude that the data presented here fairly well brackets the range of results one might expect to obtain using commercial software. The parameters of interest in this report for fingerprints are dataset quality, algorithm, and which and how many fingers are used. This report covers the dataset quality parameters and one and two-finger matches for one algorithm. Much further work is required to fully characterize all of the critical parameters of interest.

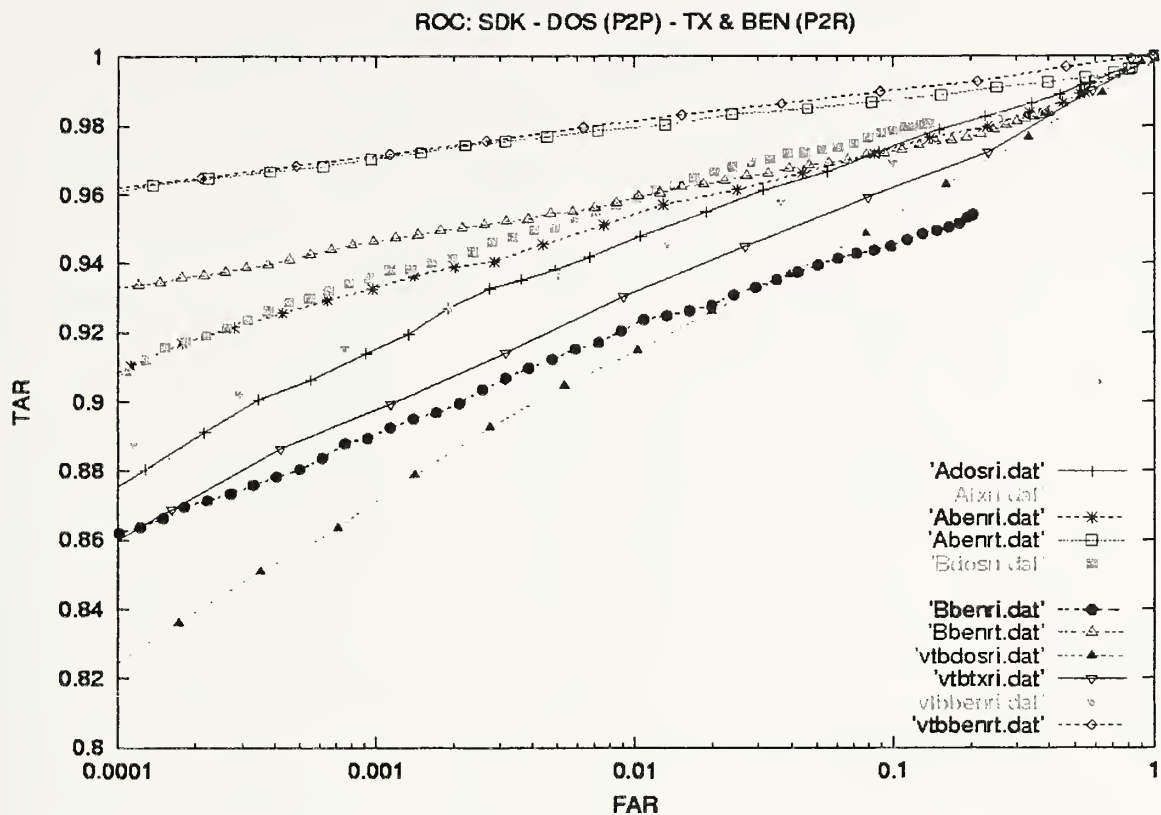


Figure 52. ROC curves for three algorithms (VTB and two commercial fingerprint matchers) and four data sets

Data Set	VTB-TAR	A-TAR	B-TAR
DOS	91.4%	94.5%	96%
TX	93%	94%	88.5%
BEN (right thumb)	98%	98%	96%
BEN (right index)	94%	95.5%	92%

Table 17. TAR for three algorithms and four data sets at 1% FAR

